Teachers of English as a Second Language of Ontario

Volume 31, No. 2
Special Issue 2005

Publication number 40038620

Special Research Symposium Issue

Acknowledgements

TESL Ontario would like to acknowledge the support it received from the following organizations and individuals to arrange for and publish the proceedings of the Fifth Annual Research Symposium:

Citizenship and Immigration Canada
Ontario Ministry of Education
TESL Ontario Conference Committee
The Participants in the Research Symposium
Teachers of English as a Second Language of Ontario

Committees

Contact Editorial
Brigid Kelso, Robert Courchêne

Reading Committee:
Hanna Cabaj, Toronto Catholic District School Board
Judith Bond, Workplace, Training & Services, Inc.
Amelia Hope, University of Ottawa
E. Marcia Johnson, Waikato University, NZ
Deborah Lee, Brock University

Table of Contents

Foreword

We are pleased to bring to you the proceedings of the Fifth Annual Research Symposium, part of the 32nd Annual TESL Ontario Conference held in Toronto in November, 2004. The themes that were the focus of the three symposia were:

Vocabulary and Technology
Teacher Knowledge - Teacher Practice
The Canadian Language Benchmarks: A Critical Evaluation

The symposia cover important issues for ESL practitioners and researchers. As subsequent language (L2) teachers, we are constantly striving to improve the quality of the programs we deliver in our classrooms. Meeting the needs of our learners, creating optimal conditions for success and frequently advocating on their behalf consumes the greatest portion of our energies, leaving little time left for reading professional journals and doing research. In organizing the Research Symposium around pertinent themes and publishing the proceedings, TESL Ontario provides its members with an opportunity to catch up on recent research and learn about their application to day-to-day teaching.

On behalf of TESL Ontario, we gratefully acknowledge the continued support of h Citizenship and Immigration Canada (Ontario Access to Settlement and Information Services) and the Ministry of Education of Ontario. Their commitment to this initiative has been an important source of encouragement in our efforts to organize the symposia and disseminate the proceedings. We hope that we can continue to count on the funding. As in previous years, the presentations included in the proceedings have again been grouped around the themes, which were Vocabulary and Technology Teacher Knowledge, Teacher Practice and The Canadian Language Benchmarks: A Critical Evaluation for 2004. These themes were selected on the basis of consultation with our membership and in conjunction with the Ontario Regional LINC Advisory Committee (ORLAC) of CIC (Ontario). They represent important instructional challenges that classroom teachers face with learners at all levels of L2 proficiency. We are confident you will enjoy reading the written presentations related to these themes and that they will provide a source of inspiration for your teaching. We would especially like to thank all the researchers who participated in the symposia and took the time from their busy schedules to submit written contributions for this publication. Without their contribution, there would be neither symposia nor proceedings to read.

Finally, we would like to thank members of the Reading Committee, the Symposia Moderators, the Conference Chair, Dorothy Coverdale, layout and formatting, the members of the Contact Editorial Committee, Renate Tilson and the TESL Ontario office staff for supporting us in organizing the symposium and in preparing the publication of this Special Research Symposium Issue. Without them and their capable assistance, our task would have been far more difficult and considerably less pleasant.

Robert Courchêne
Hedy McGarrell
Co-editors

Introduction

On the occasion of the fifth anniversary of the Research Symposium, we are pleased to bring to you this special refereed issue of *Contact* featuring papers focusing on the three themes selected for the annual TESL Conference:

The Canadian Language Benchmarks: A Critical Evaluation
Teacher Knowledge - Teacher Practice
Vocabulary and Technology

The researchers invited to presented papers examined various aspects of these themes to provide fellow researchers and language teachers with insights into the issues involved.

Theme 1: The Canadian Language Benchmarks: A Critical Evaluation

This theme was addressed in four papers focusing on task progression within the Canadian Language Benchmarks (CLB), using the CLB to teach assessment skills, to develop tests and, finally, examine their application to university and college programs. Two of the presentations have been reproduced for this special issue.

In the first paper, Janna Fox and Robert Courchêne critically examine the concept of task along with task progression within the CLB. Referring to specific examples from the various skill areas, they identity a number of problems relating to progression. To find possible solutions to these problems, they examine the current literature on tasks in L2 curriculum, SLA and L2 learning and language testing. Drawing specifically from work in language testing that focuses on task difficulty, the authors suggest a method for grading tasks -- a tentative first step in addressing issues of task progression and a possible workable approach that teachers, curriculum designers and test developers may find useful. The final section of the paper focuses on the implications of grading tasks in relation to the proficiency levels identified in the CLB.

In the second paper on this theme, Gail Stewart tackles the issues related to undertaking task-based assessment, authenticity, cultural accessibility, validity and reliability as well as how they relate to one another. After providing a brief history of the various approaches to L2 teaching, she focuses on communicative competence, drawing our attention to how it has evolved from a unitary to a multidimensional concept. On the topics of authenticity and accessibility, she questions whether using more authentic tasks that are perceived to have face validity are, by nature, valid. By using authentic tasks, do assessors replicate real-life situations, or, by definition, do all tasks and tests lack authenticity? The author cautions researchers against sacrificing validity in their desire to create authentic tasks. While it has become common to question the validity of indirect measures, direct "authentic" measures are not automatically more valid. Task design, purpose, context and interpretation of results all contribute to test validity -- tests are valid only in a given context. Addressing reliability, Stewart reminds us that validity cannot exist in the absence of reliability -- the construction of psychometrically sound tests. In terms of the CLB, she raises two important issues related to reliability: first, how

one can be empirically certain that one task is more difficult than another and second, how one resolves the potential conflict of a criterion-based instrument being used for an implicit normative function -- placing students along a continuum. She concludes this section by examining issues related to inter/intra-rater reliability writing assessment scales and how they impact on reliability. For Stewart, the increasing emphasis on high-stakes testing requires that test developers, assessors and consumers make every possible effort to ensure that tests serve the function for which they were designed.

Theme 2: Teacher Knowledge -- Teacher Practice

Three presentations focused on the knowledge base and practice of mainstream teachers (both pre-service and in-service) of ESL learners. In the sole presentation included in this volume, Tom Farrell discusses the need for ESL teachers to understand and analyze communication patterns in their classrooms. He argues that communication patterns in an ESL classroom differ from those in other settings. To underline his point, Farrell describes the nature of classroom communication and points out the various structural patterns that typically underlie such communication. Through several examples from real classroom communication, he shows how teachers might reflect on the patterns that exist in their own classrooms to gain insight into the learning process. He concludes his paper with a discussion of the implications for language teachers in managing their classroom communication.

Theme 3: Vocabulary and Technology

The theme of Vocabulary and Technology was addressed in three papers that explored web-based tools designed to help ESL learners acquire and expand vocabulary efficiently. Tom Cobb demonstrates how computers can do much more than simply download, distribute and print a broad range of materials. They can also mediate a vast repository of texts available online to enhance the amount of learning for large numbers of learners at low cost. In his paper, (available only on the TESL Ontario website), he guides ESL teachers and researchers through several features of his web site, showing how the features can enhance vocabulary development.

In the second paper on this subject, Marlise Horst presents insights from an experimental university-level ESL course in which learners used an interactive on-line vocabulary tool -- on-line concordancing -- to study vocabulary items in an academic context.

In this course, classmates used concordancers to learn new vocabulary in their academic reading. Concordancers search a large corpus of text and gather multiple examples of how a word or expression is used, allowing learners to engage in data-driven learning activities. Horst notes the various resources used as well as the students' learning gains. She contrasts high- and low users, reporting the extent to which they used these tools. The author concludes that the four case studies she presents indicate that concordancing activities may be particularly valuable in promoting the use of new words in written production.

The Canadian Language Benchmarks (CLB): A Critical Appraisal

Janna Fox, Carleton University
Robert Courchêne, University of Ottawa

> Communication is purposeful and meaningful: the primacy of purpose and meaning dictates primacy of functions over form and points to a sociolinguistic or functional theory of language. Meaning is a function of the relationships between language functions, forms and contexts, including the intentions of the speaker and the expectations of the hearer. (CLB, 1996, 12)

Introduction

Canadian Language Benchmarks—Working Paper, 1996 and now in revised form as *Canadian Language Benchmarks 2000*--have been in existence for nearly 10 years. The *Working Paper* was developed on the basis of a nationwide consultation by the National Working Group on Language Benchmarks. The mandate of this group was to establish a set of national standards that could be used as a framework for providing language training for adult immigrants. Although there had been a number of programs in the past that had attempted to meet this need, none had resulted in the creation of a set of standards that could be used by all stakeholders across the country. In accordance with Canada's bilingual and bicultural heritage, such standards were to be developed for both English and French.[1]

Paralleling the appearance of the CLB in 1996 was the creation of The Centre for Canadian Language Benchmarks, whose mandate was to oversee the Benchmarks project. Since its inception, the CCLB has commissioned projects that have resulted in the development of assessment instruments, (Canadian Language Benchmarks Assessment (CLBA) and the CLB Streamlined Placement Test), curriculum guidelines, pedagogical materials as well as other support materials and implementation guides.[2] Its ongoing support of the CLB has resulted in their gradual acceptance as a recognized standard across the country.

Along with acceptance of the CLB has come a closer examination of their theoretical underpinnings and a questioning of the suitability of their descriptions of language proficiency (competencies, sample tasks) for all levels and contexts of L2 teaching and learning, including college and university. As a result of consultations in the field, the CLB were revised in 2000 to incorporate a number of suggested changes; specifically, the separation of listening and speaking in both descriptive and assessment terms. Such changes made it possible to establish equivalences between the CLBA and other international tests (e.g., TOEFL, IELTS, CAEL, CELPIP, CanTEST) accepted in Canada for university admission and professional certification. Such tests provided candidates with a separate score for listening and speaking. In the ongoing revision of the CLB, one element that has not received critical scrutiny is the concept of *task,* despite frequent

[1] The *Standards linguistiques canadiens* were published in 2002.
[2] For a complete list of materials available visit www.language.ca

requests from the field for clarification regarding task definition and progression (i.e., scope and sequence).Tasks are central to the CLB as they are designed for the dual purposes of: 1) providing teachers and curriculum designers with an indication of what learners can do at increasing levels of L2 mastery; and 2) providing teachers and test developers with scale criteria for placement and assessment purposes. Indeed, both the CLBA and the CLB Streamlined Placement Test are task-based instruments.

In this paper, we begin by examining the concept of task as used in the CLB and the CLBA, relate this analysis to the current literature on tasks, drawing specifically on research available from the disciplines of L2 learning and testing and, finally develop a series of guidelines for analyzing task difficulty in light of current research.

The Concept of Task in the CLB

According to the developers, the CLB are 'descriptive scales of communicative proficiency in English as a Second Language (ESL) expressed as 12 benchmarks or reference points. (Pawlikowska-Smith, 2000, viii). They describe what a learner should be able to do with language at a given benchmark level under a specific set of performance conditions (i.e., length of text, time constraints, audience). They also provide sample tasks as well as outcomes to determine if the learner has succeeded in performing the task and/or achieved a given benchmark level. In addition, they function as a set of standards that can be used by curriculum designers, language teachers and assessors to produce materials and tests. They are intended for use by government, industry and post-secondary institutions as measures of a candidate's L2 proficiency. Finally, with this emphasis on tasks, they place the learner and what (s)he can do with the language at the centre of the learning process, emphasizing that being able to *do* is more important than *knowing how* to do, as indicated in the following:

> A Canadian Language Benchmark is a description of a person's ability to use the English language to accomplish a set of tasks. (Ibid, ix).

Why this emphasis on the importance of tasks? First, the underlying view of language assessment in the CLB is criterion referenced. Tasks, which are 'directly observable and measurable performance outcomes' allow one to evaluate a learner's ability to 'perform' at a certain competence level. The CLBA, used to determine learners' benchmark levels in the four skill areas, consists of a series of tasks measured against a given set of criteria. Second, tasks parallel what people do in real life with language. Consequently, the proficiency descriptions in the CLB have a language competence focus as opposed to a linguistic one. The emphasis is on proficiency rather than grammatical mastery, the latter not being directly tested in the CLBA. Third, an underlying assumption is that tasks can be manipulated in terms of complexity, performance conditions and outcomes to become increasingly difficult as one progresses from benchmark 1- benchmark 12 (BK). Four, another underlying assumption is that the four language skills, the three stages (Basic, Intermediate and Advanced), the four competency areas (social interaction, following and giving instructions, suasion and information) along with the three contexts (Community, Study and Workplace) can all be described in terms of task difficulty and progression.

Finally, according to the CLB, tasks promote the integration of all the four skill areas along with the acquisition of different aspects of communicative competence.

Thus, the definition of task adopted by the authors of the CLB (though not explicitly stated) is essentially that proposed by David Nunan (1989).

> A communicative task is 'a piece of classroom work which involves learners in comprehending, manipulating, producing, or interacting in the target language while the attention is principally focused on meaning rather than form. The task should also have a sense of completeness, being able to stand alone as a communicative act in its own right'. (p.10)

In terms of the distinction made in the field between 'focused tasks' (Long and Robinson (1998) use the term, 'focus-on-form'; Loschky and Bley-Vroman, 1998) and 'unfocussed tasks' (Ellis, 2003; Skehan, 2003), the position adopted in the CLB is that of the unfocused task where the emphasis is on communication. While a given task may predispose a learner to choose a specific form, there is no obligation to do so, while in focused tasks the intention is to draw the learner's attention to a specific form and/or have him/her use a specific form.

In the introduction to the CLB-Working Paper, 1996, the authors go to great lengths to explain that many factors must be taken into consideration in determining the difficulty of a specific task (Table 1):

-directedness to the immediate context	-speed
-factuality	-status of the interlocutor
-predictability	-decision making involved
-length	-purposes and consequence of communication
-subject matter	-amount of detail
-accuracy/precision involved	

(Ibid, 15)

Table 1: Factors Affecting Task Difficulty in the CLB

For example, in discussing reading tasks, they comment:

> The choice of a specific reading text for a specific task to be accomplished at a certain proficiency level can be guided by general guidelines of text characteristics but, in the end, must be pragmatically established by exposing intended users to the text and the task, and observing and estimating their levels of performance (1996, 16).

According to the authors, while certain characteristics inherent in a given task can influence its difficulty, it is the performance and situational conditions that determine the level of performance required to carry it out. As well, issues of 'ease versus difficulty' and 'simplicity vs. complexity of language' are also influenced by many factors that are

frequently unpredictable. In this sense, tasks are not any different from functions, notions, structures or skills in that the criteria for setting up a logical and increasingly difficult progression are hard to determine (Brindley, 1989; Candlin, 1987; Nunan: 1989, Skehan, 1998; Robinson, 2001).

Ellis (2003), in his comprehensive review and synthesis of the literature on tasks, has identified the following criteria for grading tasks.

Criterion	Easy	Difficult
A. Input Medium Code complexity Cognitive complexity -information type -amount of information -degree of structure -context dependency Familiarity of information	pictoral → written High-frequency vocabulary; short and simple sentences -static → dynamic -few elements/relationships -well defined structure -here-and-now orientation -familiar	→ oral low-frequency vocabulary; complex sentence structure → abstract -many elements/relationships -little structure -there-and-then orientation -unfamiliar
B. Conditions -Interactant relationship (negotiation of meaning) -Task demands -Discourse mode required to perform the task	-two-way -single task -dialogic	-one-way -dual task -monologic
C. Processes -Cognitive operations -type -Reasoning need	-exchanging information -reasoning: few steps involved	→ exchanging opinions - many steps involved
D. Outcomes -Medium -Scope -Discourse of task outcome	-pictorial -closed? -lists, descriptions, narratives, classifications	→written →oral -oral? → instructions, arguments

(2003, 228)

Table 2: Criteria for Grading Tasks

A close examination of the criteria found in Tables 1 and 2 reveals a considerable overlap, providing evidence of theoretical resonance. However, theory does not guarantee the resolution of practical problems. Therefore, in the next sections, we address the practical problems of task progression within the CLB which create challenges for teachers, testers and curriculum designers in their practice.

Task Progression within the CLB

In this paper, it is not possible to examine in detail the task progression across the four skill areas. We have, therefore, chosen to focus primarily, not exclusively, on writing.

In the Table reproduced below (Table 3) from the CLB, we have isolated the types of writing texts that L2 learners should be able to produce at different Benchmark levels.

Benchmark	**Writing: Task Descriptions**
1	Can write a small number of familiar words, simple phrases and sentences about self related to immediate needs.
2	Learner can write a few sentences and phrases about self and family or other highly familiar information as a simple description, as answers to written questions or on simplified forms and slips.
3	Learner demonstrates adequate competence in simple, familiar, personal writing tasks within predictable contexts of everyday needs and experience. Can write a number of one-clause sentences about self and the family (e.g., simple descriptions and narration)
4	Learner can effectively convey in writing simple ideas and information about personal experience within predictable contexts of everyday needs. Can write short descriptions and narration of events, stories, future plans abut self and family, or other highly familiar topics.
5	Can effectively convey an idea, opinion, feeling, or experience in a simple paragraph.
6	Can write one- or two-paragraph letters and compositions.
7	Can join two or three paragraphs into larger text.
8	Can link sentences and paragraphs (three or four) to form coherent texts to express ideas on familiar abstract topics, with some support for main ideas and an appropriate sense of audience.
9	Can write a coherent essay, paper, or report (three to five double-spaced typed pages; descriptive, narrative, expository, argumentative/persuasive) in order to present information and state a position on a previously researched topic.
10	Can write an effective stylistically complex and interesting essay, story, paper or report (10 double-spaced pages) about a previously researched topic.
11	Can write an effective, stylistically complex and interesting sizable text (up to 20 double-space pages); expository or argumentative essay; symbolic or allegorical story; rational inquiry paper; problem-solution paper; or analytic report about a previously researched topic.
12	Can produce effective and stylistically polished essays, documents, articles, theses (over 20 double-spaced typed pages or other volume appropriate to purpose and audience).

(CLB 2000, 40, 98, 168)

Table 3: Selected Writing Task: Benchmarks 1-12

Length

While it is pointed out in the CLB that length does not equal difficulty, there is a very close correlation between the length of text and Benchmark level across the four skill areas. In the above table, as one progresses from Benchmark 1 to Benchmark 8, the

length of text to be written increases on an incremental basis until it reaches three-to-four paragraphs at Benchmark 8. We find such a progression worrisome for a number of reasons. First, on the basis of research done to date, Benchmark 8 is considered the level students need to attain in order to undertake full-time undergraduate programs at universities or colleges. Considering that a candidate assessed at Benchmark 8 has met some but not all of the requirements for this benchmark, a text of three-to-four paragraphs in length is certainly not what is required in terms of writing skills to undertake full-time study.

Second, one would expect that from one benchmark level to another, considered to be about 250 hours of instruction, a candidate would certainly make more progress than going from a one-paragraph to a two-paragraph text in 1000 hours. Third, in examining the writing performance required at Benchmark 9 as opposed to 8, one can only wonder at the incredible jump in task-difficulty level. How does one go from producing a three-to-four paragraph text to writing coherent 4-5 page essays involving a number of different rhetorical patterns in 250 hours? After this leap from Benchmark 8 to 9, length, again, assumes an important role as one moves from one benchmark level to another.

If we look at the performance descriptors in reading for Benchmarks 8 and 9 (see Table 4 below), we find similar concerns regarding length. In these descriptors, however, no explanation is provided as to why someone, who, supposedly, has the proficiency to do full-time undergraduate/graduate work, has to read texts of just three-to-five pages in order to meet the level requirement. From the beginning of a post-secondary program, students are typically asked to read lengthy textbook chapters, research articles, complex reports, novels, monographs, etc., that are clearly much longer and more difficult.

Benchmark	Reading: Global Performance Descriptors
8	Learner can follow main idea, key words and important details in an authentic two-to three-page text on a familiar topic, but within an only partially predictable context. Texts can be on abstract, conceptual or technical topics, containing facts, attitudes and opinions. Inference may be required to identify writer's bias and the purpose/function of the text.
9	Learner can read authentic multipurpose texts, daily newspaper items, short stories and popular novels; academic materials, sections of textbooks, manuals; simple routine business letters and documents. Texts are three to five pages long, proportionally and linguistically complex, but with clear underlying structure, on abstract, conceptual or technical topics.

Table 4: Benchmark 8 and 9 Global Performance Descriptors

Again, one is struck by the incongruity between the formal performance descriptor for Benchmark 9, namely that, students must be able to read texts that "are three to five pages long" and the actual performance demands that learners at this level face; e.g., the average graduate student needs to read approximately 200 pages a week. Although the CLB is neither a curriculum nor a test according to its developers, providing details

regarding text length and sample tasks leads anyone using the document to use these as guidelines for task development. Further, it is not surprising that curriculum designers, language teachers and assessment experts have struggled to find the underlying *spiral of difficulty* in the CLB criteria and tasks (Witol, 2003).

Issues of acculturation

In enumerating the different factors that can affect task difficulty, an important omission in the CLB is the relationship between learner characteristics and task type (Skehan, 2003). In interviews, one of the co-authors conducted with students at La Magie des Lettres, an adult education centre, to evaluate their ability to understand and speak French, it soon became evident that there was a method effect: for students who are not literate in their own language (literacy here defined as being able to read and write) and who had no academic training, describing a series of pictures, listening to a text and summarizing it were activities that they had never done. Consequently, they were not able to provide the best sample of their ability in French. These candidates possessed Basic Interpersonal Communication Skills (see Cummins' concept of BICS, 1984, 1996) to varying degrees but the tasks used did not allow them to demonstrate it. While the CLB criteria for task difficulty are certainly pertinent, they represent only one side of the interaction--tasks are enacted in contexts by real learners who bring to the task various sets of characteristics that constrain or enhance their ability to accomplish them. We have listed some of these characteristics in Table 5 below.

Learner Variables	**Cultural Variables**
Age	Role of the teacher and learner
Education	Interpersonal relationships
Literacy level	Literacy and orality orientations
Learning styles	Rhetorical patterns
Gender	Views of gender: power relationships
L1	
Previous language experience	N.B. These criteria are not mentioned
Views of language learning	in the CLB

(Courchêne and Fox, 2004)

Table 5: Some Learner and Cultural Variables Affecting Test Performance

Related to this is the fact that a student's construction of their identity (i.e., their cultural construct), interacts with how they perceive and learn language, and by corollary, how they perform in assessment contexts (Abedi, 2003). While we are not suggesting a strong version of the Sapir-Whorf hypothesis, (the notion that how we categorize and organize the world is determined by our native language), students' cultural constructs certainly influence how they speak and write, or "do" language. There is no recognition in the CLB of the influence of learner and cultural variables in task performance as they relate to increasing levels of L2 proficiency.

Task difficulty

The CLB identify three proficiency stages (basic, intermediate and advanced), with four benchmarked standards at each stage. According to the CLB scale, these stages are marked by "progressively more demanding communication tasks and contexts, and progressively higher expectations and quality of communication" (CLB, 2003, xi). However, although the CLB descriptors of proficiency standards provide a general overview of increasing language proficiency, individual institutions, program coordinators, curriculum developers and language teachers across Canada have been left to their own devices in managing the inherent ambiguity in the proficiency descriptors themselves, as well as the messiness that characterizes the progression of task difficulty. It is important to note that setting workable parameters for task difficulty has long been a concern in research within language curriculum, second language acquisition (SLA), second language (L2) learning, and language testing. Within these research literatures, the challenges with current approaches to investigating task difficulty have been widely discussed at both conceptual and methodological levels (see particularly, Bachman, 2002; Bygate, Skehan and Swain, 2001). Although there may be evident differences in the approaches taken by different researchers both within and across these disciplines in Applied Language Studies, it may be useful and informative to examine what these different research literatures have contributed to our general understanding of task difficulty, and to consider how this may assist language teaching professionals in setting the scope and sequence of tasks for learners in CLB-related programs. The following section of this paper provides a selective but representative overview of some key research relating to task difficulty and sequencing in language curriculum, second language acquisition (SLA), second language (L2) learning, and language testing.

How is task difficulty addressed in English Language Teaching (ELT) curriculua?

The sequencing of tasks and definitions of their difficulty have been an on-going challenge in task-based curricula (Breen, 1987; Brown, 1996; Norris, Brown, Hudson, Yoshioka, 1998; Prabhu, 1987; Short, 1993). Different curriculum theorists have identified different means of addressing this issue through, for example, systematic needs analysis (Brown, 1996), rich discursive and/or ethnographic accounts of contexts of use (Breen, 1987; Swales, 1987), the definition of task parameters in relation to theoretically-defined frameworks (Bailey, 1998; Prabhu, 1987), and even trial and error (Hajer, 2003). One of the earliest and perhaps most extensive elaborations of a task-based curriculum may be found in the work of Breen (1987). Breen defined task-based curricular and syllabus approaches as ones that "organize and present what is to be achieved through teaching and learning in terms of how a learner may engage his or her communicative competence in undertaking a range of tasks" (p. 160). Breen argued that task-based approaches simultaneously address not only *how to learn* but also *how to communicate*.

He drew a distinction between what he refers to as 1) communicative tasks and 2) *meta*-communicative or learning tasks. Whereas the communicative task focuses upon the actual sharing of meaning through communication in relation to a specific purpose; learning tasks focus on "exploration of the workings of the knowledge systems

themselves and, in particular, how these may be worked upon and learned" (p. 161). Breen suggests that these two types of tasks should be derived from a systematic and thorough collection and examination of tasks at play in the context of target language use (TLU). Once tasks have been collected, they should be analyzed for the underlying system of rules, conventions, patterns of language skill and background knowledge that are required in order for a learner to participate through language with accuracy, appropriateness, and effectiveness in managing each task within the TLU context. Breen suggested that *frequency* and *relevance* should guide the selection of communicative tasks: "the designer will select and cluster those tasks for the syllabus that are most common in the target situation or most relevant in terms of learner need and interest, or through some combination of these selection criteria" (p. 162). Further, Breen discussed the cyclic sequencing of tasks in relation to difficulties experienced by learners in engaging with them. Thus, according to Breen, enactment of a task-based curriculum involves on-going or dynamic monitoring, assessment and planning in relation to individual teachers' classroom-based observations of learner needs and responses.

The sequence in which tasks are introduced, according to Breen, should relate to two fundamental criteria:

- learners' familiarity with a task, and
- task complexity (i.e., how much is demanded of learners).

It is evident that the demands placed on teachers to define task scope and sequence, however, are considerable in Breen's model, not to mention the requirements for research-intensive specification of the TLU context(s) and analysis of tasks for their underlying language dimensions. Within the Breen model, the quality of the curriculum is directly dependant on how comprehensive and accurate the research is that defined it. Nonetheless, the focus on classroom-based, teacher decision-making within this dynamic/responsive model is both this model's greatest strength and weakness. On the one hand, it is learner-centred and directed; on the other hand, because it is so individualized and embedded within unique learner-teacher interactions which are dependent upon teacher "noticing" and awareness of what should happen next, it may be impossible to generalize from one classroom to another, or to allow for comparative evaluation of the systematic effectiveness of the curriculum and its activities in relation to overall learner needs and goals. This has been one of the principle criticisms of task-based curricula. Ironically, tying the CLB criteria to tasks within the Breen model potentially introduces the same levels of idiosyncratic variability that the Benchmarks were designed to eliminate. Over the years, the variability inherent in the Breen model has been addressed in a number of ways by different researchers.

Crookes (1986) and Long (1985), for example, identify a number of hypothesized factors in order to provide a fixed framework for evaluating task difficulty, including: the number of steps required, the number of participants involved, the knowledge required to complete the task, the degree of challenge and space/time constraints. Nunan (1989) identifies other factors to consider, such as the amount and quality of input provided, as well as learner factors (e.g., confidence, prior learning experience, linguistic knowledge,

motivation) and activity factors (e.g., complexity, help available to the learner, communicative stress, the amount of context provided for the task). Nunan's framework for the analysis of task difficulty is derived from classroom-based observations, but these observations were not actually grounded in empirical research, and there is little empirical evidence of the use of Nunan's framework in practice.

A recent example of a curriculum that is operationalized by tasks is the *Target-Oriented Curriculum* (TOK), instituted by the Education Department of the Hong Kong Special Administrative Region Government, which has been implemented in the primary schools in Hong Kong. In conception and task definition it greatly resembles the CLB. Candlin (2001) provides a helpful distinction, however, between this target-oriented curriculum and task-oriented curriculum. The "targets" in the case of the TOK are goals or aims regarding the competencies, skills and strategies that learners should develop; the "tasks" are the *means* by which learners would be involved in using language in relation to the curricular "targets". Candlin (2001) notes that one might argue for task sequence in relation to any of four task components identified in the curriculum, namely, purpose(s), processes, outcome(s) and modes of evaluation (p. 241).

As Candlin points out, "what the TOC in Hong Kong does exemplify is an intense awareness of the curriculum planners of the recent literature in language acquisition and pedagogy and a strong willingness to see a generally held current view of language as communication and of language learning as process, and the classroom as an interactive site of engagement, to permeate its curriculum guidelines" (p. 237). The key point of Candlin's explanation of the TOC is that tasks are viewed as the means by which the curriculum targets are achieved – rather than outcomes in themselves.

Although Candlin identifies three "components" to consider in deploying tasks within the curriculum, namely, *task design*, *operationalization* and *evaluation* (p.238), he points out that these components are only tentative attempts to deal with task dimensions in practical terms, and that neither the TOK nor the research literature has managed to identify *how* tasks might be sequenced in relation to the targets of the TOC. Candlin does not seem to consider this a problem, even though he recognizes that "central to undertaking the design and operationalization of any public curriculum in the sense of a planned, institutionally-based program, is the need to select, to grade and to sequence." (p. 240). The task components he identifies in this regard could potentially inform task selection and sequence, but the TOC does not address the need for sequence directly or provide any systematic support for teachers in selecting and ordering tasks in their implementation of the curriculum. Thus, the same ambiguity that has plagued the relationship between criteria/descriptors and tasks in the CLB is evident in Candlin's analysis of the TOC, when he concludes that,

> Despite an increasing amount of empirical studies in the development of interlanguage pragmatics (Kasper and Rose, 1999), we may ask why has there been little *curriculum-oriented* research which seeks to connect task design and operationalization with the systematic development of learner discursive strategies and pragmatic behaviour. (p. 241)

He suggests that second language acquisition and learning theory research might be an important source of information with regard to task hierarchy, difficulty and sequence. And, there is, in fact some important work being done in this area.

What can Second Language Acquisition (SLA) and Learning Theory contribute?

A number of researchers (see, for example, Swain, 1995; Swain and Lapkin, 2001; Samuda, 2001; Samuda and Rounds, 1993), building on years of research focusing on features of *interlanguage* (IL), have begun to explore, explain and focus attention on issues in task design in both research and teaching. It may be, as Samuda (2001) argues, that the ambiguity with regard to task difficulty, sequence and selection and the centrality of tasks in curriculum and pedagogical approaches, have lead to a general problem in the use of tasks in language teaching, namely,

> The use of tasks has long been recognized as a central feature of communicative language teaching, and as such is widely addressed in teacher education, and in the development of instructional materials; so much so that, for many teachers primed to deal with tasks for activating language and stretching fluency, task based language teaching (TBLT) has become synonymous (albeit misguidedly) with unscripted oral activity. (p. 120)

Thus, issues that have been raised by researchers and theorists examining task-based curriculum, have also become a focus of research in Second Language Acquisition (SLA) and learning as well. Of central concern is the teacher's role in acquisition and learning. As Samuda (2001) points out, in task-based language teaching (TBLT) the relationship between task and teacher is essentially "complimentary" (p. 120). However, Samuda's research demonstrates that task design itself can potentially influence how teachers and students interact with tasks. She identifies a distinction between two types of tasks that is similar in many respects to that identified by Breen, namely:

1) tasks designed to activate, stretch and refine current interlanguage (IL) resources and processing capacities (what Samuda refers to as "language-activating/fluency-stretching tasks") and
2) tasks that help learners make new form-meaning connections (referred to by Samuda as "knowledge-constructing tasks").

Samuda's (2001) research demonstrates that task design and the management of attentional focus is a means of encouraging language development. Samuda bases her task design on a *meaning* → *form* → *meaning* connection:

> The task design explored here initially seeks to create a semantic space; as learners orient themselves within this space, they may be pushed by operations carried out on the task input data to notice holes (Swain, 1995) in their current IL resources. In the potential learning space thus formed, opportunities may be created for new form-meaning mappings to be made. The nature of the

> highlighted form meaning relationships may also affect initial design options, and determine where it is more appropriate for the task input data to be linguistically seeded or semantically enhanced. (p. 124)

What Samuda's research illustrates is not only that task design can direct the attention of learners to particular meaning-form connections, but also that teachers play a critical role in supporting those connections. She concludes that tasks provide an important source of input to be "mined" (p. 137) by both teachers and learners in supporting learning. Swain and Lapkin (1998, 2001) have also designed tasks to encourage talk and direct learner attention to "noticing" (Schmidt, 1990) important meaning-form connections. In their work with tasks such as the dictogloss task and the jigsaw task, they have identified salient differences in learner generated *language-related episodes* (Swain and Lapkin, 2001, p. 131), which they define as discourse that arises in the flow of meaning-based talk, in which learners focus their attention on form. They find that whereas the dictogloss task focuses learner attention on accuracy and constrains the range of students' time on task, the jigsaw task "led to a greater range of vocabulary use and language-related episodes, suggesting that perhaps its open-ended nature might inspire greater linguistic creativity" (p. 111). Understanding differences in learner-talk that are generated in response to specific tasks is helpful and useful information. However, it does not help with the issues of difficulty, sequence or scope.

In a recent interview, Swain (cited in Fox, 2004) recounts how research that investigated "noticing" across different proficiency levels demonstrated that lower proficiency students noticed less. When lower proficiency students were allowed to use their first language, their quality of noticing increased. Swain suggests that only students at a certain level of proficiency are able to use information that they learn in a co-constructed language related episode and apply it on their own in a subsequent task. This type of "transfer" may be accounted for by a threshold. The identification of such thresholds would suggest a promising hierarchy of performance that could be tied to benchmark criteria and linked directly to task design. Although task design issues have received increasing attention from L2 acquisition and learning researchers, their research has only begun to document the implications. The issues of task difficulty have been most directly addressed within the field of language testing where, with the increasing use of task-based performance assessments, both the hierarchy of task difficulty and the comparability of tasks across multiple versions of a test have posed a considerable threat to both arguments for validity and reliability. It is within the field of language testing that the most dramatic progress has been made in characterizing task difficulty. In the next section, one of the more promising approaches is described in some detail. It may, in fact, provide to curriculum developers, teachers and assessors who are working with the CLB a systematic and empirically grounded means of grading task difficulty. It is not without controversy, however.

How can language testing help?

Bachman (2002) discusses the challenges that have been faced by language testers in the identification (or specification) of task difficulty and warns against what he refers to as

task-driven assessment. He argues that the proper focus of testing is *construct-driven assessment*. The term "construct" refers to *what* testers attempt to measure – be it a skill, ability, or performance within a specific context. Testers attempt to elicit behaviours on tests that are representative of what is being tested. When tasks themselves become confused with the construct of interest, however, there is a risk that insufficient behaviours are elicited by the test and the underlying construct is ultimately underrepresented. As Messick (1994) has convincingly argued, construct under-representation undermines the validity of inferences drawn from test performance. In other words, it renders the testing process meaningless.

Bachman takes issue with some task-based test development for this reason – charging that tasks themselves are poor and misleading substitutes for constructs of interest. He is critical of the work of Brown, Hudson, Norris and Bonk (2002) in spite of its rich theoretical ground in Skehan's (1996) research. Skehan identifies features of task difficulty that are both theoretically and empirically derived and defines the ability requirements and task characteristics inherent in any given L2 task (p. 50). It is Skehan's framework that grounds the research of Brown et al. (2002) and provides a practical and useful approach to task difficulty.

As Brown has argued, task-based performance tests were different from other types of tests in that students taking them were asked to use language in order to "perform tasks much like what would be expected of them in the real-life situations in which they would eventually be expected to use the language" (Brown, 1996, cited in Norris et al., p. 8). The focus on "performance" (i.e., observable language behaviour) elicited by tasks that were representative of tasks encountered in "real-life" necessitated not only considerable work on scale development and the elaboration of criteria for scoring and interpreting test behaviours (many of which were task-specific), but also more sophisticated methods of rater training and rater monitoring in order to control for reliability. Most importantly in relation to the CLB tasks, Brown et al. drew on the Skehan framework, in elaborating a methodology for grading task difficulty, allowing for both the development of a task-based language curriculum (comprised of tasks that were systematically controlled for difficulty), and also the development of task-based assessment – linked to and reflecting the curriculum. Although it is instructive to keep Bachman's concerns regarding the dangers of task-driven assessment in mind, an overview of the Brown et al. framework is provided below. (See Table 5). It may be very helpful to teachers as a means of grading and sequencing tasks in relation to the CLB proficiency criteria.[3]

[3] There are many other researchers that have studied factors influencing task difficulty. The review presented in this paper is selective. Teachers, curriculum designers and researchers may wish to consult other studies for useful information regarding other factors, including the influence of a group in supporting task performance (Fulcher, 1996), the role of linguistic features of tasks (Robinson, 1995) and other factors (e.g., Crookes, 1989; Foster & Skehan, 1996; Yule & Powers, 1994). These studies also utilize a range of research approaches that may be useful and informative, as well, including observational techniques, retrospective, post-task interviews, unique classroom-based research designs, etc.

	Easy → Difficult	Easy → Difficult	Easy → Difficult
	Range	*Number of input sources*	*Delivery of input*
Code	- +	- +	- +
	Amount of information to be processed	*Organization of input*	*Availability of input*
Cognitive complexity	- +	- +	- +
	Mode	*Channel*	*Response level*
Communicative demand	- +	- +	- +

Table 5: Language Performance Task Difficulty Matrix (Brown, Hudson, Norris & Bonk, 1998, p. 77)

Each one of the three criteria identified in Table 5 above, namely (1) code, (2) cognitive complexity and (3) communicative demand, is applied by Brown et al. in grading a task across the nine characteristics that define the criteria: (*criteria 1/code*) range, number of input sources and delivery of input; (*criteria 2/code complexity*) amount of information to be processed, organization of input, and availability of input; (*criteria 3/communicative demand*) mode, channel and response level. Although Brown et. al (2002) continue to research and revise their matrix, the original matrix and categories are presented here. Grading occurs across nine dimensions in the matrix and involves a yes/no decision in each cell with "a minus sign always [indicating] less difficulty with respect to the [criteria] and characteristic relative to the given task, whereas a plus sign always indicates greater relative difficulty"(p.77). Tasks are graded by counting the number of +'s that are awarded. Thus, tasks can range in difficulty from 0 to 9, with 0 indicating the task is *very easy* and 9 indicating the task is *very difficult*. In the section below, further explanation is provided for each of the criteria and characteristics in the matrix.

Criteria and Characteristics

Code: What kind of language and information is involved in successful task performance?
How difficult are the vocabulary, grammar, text structure, pragmatics and pronunciation involved in the information that must be processed and produced in order to complete the task?

Characteristics (Three features are considered in relation to code)

1) ***Range*** *(Is the language restricted (i.e., basic vocabulary, simplified grammar or authentically consistent with text/discourse one would find in actual contexts of use?)*
2) ***Number of input sources*** *(Does the performance require the learner to process multiple sources of input in order to accomplish the task? Are there two or more independent sources of information that must be processed in order to complete the task?)*
3) ***Delivery of input*** *(How accessible is the input? Is the quality of the input simple and clear? Or, is the input complex (e.g., multiple differing accents, mixed registers, rapid, layered?)*

Cognitive Complexity: Does the task appear to create a considerable cognitive load?

Characteristics (Three features are considered in relation to Cognitive Complexity)

1) ***Amount of information to be processed*** *(Is the learner required to use a considerable amount of information in order to complete the task?)*
2) ***Organization of input*** *(Does the performance require the learner to significantly re-organize the input in order to complete the task?)*
3) ***Availability of input*** *(Does the learner need to search for and select information from multiple sources, make decisions in the course of the task about the relevance of the information provided and/or gaps in the information, identify alternative resources?)*

Communicative Demand: Does task require complex communicative activity?

Characteristics (Three features are considered in relation to Communicative Demand)

1) ***Mode*** *(Does the task require extensive productive communication in speaking and/or writing in order to complete?)*
2) ***Channel*** *(Does response require that the learner process information from multiple sources involving both auditory and visual processing (e.g., annotating a written outline for a lecture delivered in real-time with supporting visual diagrams?)*
3) ***Response level*** *(Does the task require immediate or "real-time" completion?)*

Brown et al. (1998), following the advice of Long and Crookes (1993), provide multiple examples of the design and grading of tasks in relation to themes relevant to a particular

context of use (i.e., student interest, purpose for studying English, etc.). Bachman would take exception with other advice of Long and Norris (cited in Brown et al.1998), who argue that the "construct of interest in task-based assessment is performance of the task itself" (p. 9). It is important to keep in mind, when reviewing these examples, that the *targets* (Candlin, 2001) or CLB criteria –are the "constructs" of interest. The tasks are a means of eliciting language use in relation to the target constructs.

Below is sample task from the theme "At the airport" – a collection of tasks that stress receptive comprehension skills and specific information. The task below would be supported by realia such as a ticket with information about destination and flight number.

Task: Find your flight

Prompt:
At the airport you want to find out where and when your flight is boarding. Look at the flight departure screens and try to identify your flight. Note the salient information regarding your flight from the departure screen.

(Source: Norris, Brown, Hudson, Yoshioka, 1998, p. 212)

With regard to code, the difficulty of the task would be restricted to low, which Norris et al. describe as, "Simple understanding of what constitutes salient information in the situation." (p. 212). As well, the cognitive complexity is very low, as processing involves simply matching key information. The communicative demand is also very low – only accuracy is required to successfully complete the task. In grading this task, Norris et al. give this task a 0 overall, assigning a (–) to each of the characteristics in the matrix[4].

How could the matrix be applied to the grading of a task taken from the CLB? An example was drawn from CLB level three/four which is defined by the following criteria:

> Simple structure is mastered at this level. Messages are short and are limited to very basic daily routine situations. There are frequent errors and often a need for clarification and repetition. Can read and write very short simple texts with recognizable spelling and punctuation. Vocabulary is quite limited.

Suggested tasks at these levels include activities such as filling in a form or taking down information from a phone message. Below is an example of a task developed in relation to this CLB criterion. Subsequently, the task is graded for difficulty using the Norris et al. matrix.

[4] In this example, Norris et al. apply a matrix with just six characteristics – two for each criteria – rather than the full matrix included in Table 1.

Sample CLB Task: Filling-in a form

Prompt: You need to provide the following information in order to receive a free magazine subscription that is required reading for a course that you are taking. When you have finished filling in the form, please give it to the instructor.

Research Record Chronicle

Dear New Subscriber: We are offering you three free months (3 issues of *Research Record Chronicle*) in exchange for a little extra information about you. Required information is marked (R), extra information is marked (E). If you do not provide all of the information requested by the form, please remit $55.00 with this form.

(R) Name _____ _____ (E) Sex F M
Last First Initial

(R) Address _____ _____ _____
Number Street Suite

(R) _____ _____
City Province

(E) Age: Below 20 ☐ 20-25 26 ☐ 26-35 36 ☐ 36-45 46 ☐ 46-55 ☐ Above 56 ☐

(R) _____ _____ (E) e-mail _____
Area Code Telephone number

(E) Marital Status _____ (E) Occupation _____

(E) How often do you read the following in approximate number of hours a week?

	Not at all	1 hour or less	2-7 hours	more than 8 hours
English Newspapers	☐	☐	☐	☐
English Magazines	☐	☐	☐	☐
English Best Sellers	☐	☐	☐	☐
English Academic Texts	☐	☐	☐	☐

(E) How many hours a week do you watch television in English? _____

(E) If you watch television, name two of your favourite programs:

1) _____

2) _____

(E) If you read English newspapers, which is your favourite?

_____

(E) If you read English magazines, which is your favourite?

_____

(E) If you read best sellers, name a favourite:

_____

Language Performance Task Difficulty
Grading the Sample CLB Task

	Easy → Difficult	**Easy → Difficult**	**Easy → Difficult**
	Range	*Number of input sources*	*Delivery of input*
Code	**- (1) +**	**- +**	**- +**
	Amount of information to be processed	*Organization of input*	*Availability of input*
Cognitive complexity	**- (1) +**	**- +**	**- (1) +**
	Mode	*Channel*	*Response level*
Communicative demand	**- +**	**- (1) +**	**- +**

Example: Grade = 4

Summary

Code	**Range +**	**Number of input sources -**	**Delivery -**
Cognitive Complexity	**Amount of info. +**	**Organization -**	**Availability +**
Communicative demand	**Mode -**	**Channel +**	**Response –**

When considering tasks for grading, it is useful to have a number of colleagues grade them independently, and then, discuss as a group the reasons why a plus (+) or a minus (-) was assigned to each of the task characteristics. It is particularly important to discuss differences in grading. Over time, consistency in grading will develop as the rationale for assigning a (+) or (–) in each cell of the matrix becomes a matter of well-articulated practice.

This approach offers a systematic method of grading task difficulty. Applying this approach does not suggest that this is *task-driven* curriculum or assessment, however, because the tasks are consistently related to the CLB criteria descriptors. It is the criteria that are the constructs of interest. Therefore, the approach is arguably construct-driven.

One of the other direct benefits of the approach is that it necessitates discussion among teachers, curriculum developers or testers about the features that characterize differing levels of proficiency, as described by the CLB. Not only will such discussion potentially increase consistency within and across levels, but it will also lead to task benchmarks – multiple tasks that are indicative of key CLB proficiency levels within and across

programs. The grading and sequencing of tasks according to a principled and systematic methodology will lead to greater task comparability and consistency in relation to the CLB criteria. Currently, the CLB simply identify criteria and sample tasks but do not provide a method for sequencing or structuring tasks within an empirically or theoretically derived framework of progression. The work of Brown et al. (2002) and Norris et al. (1998) provides a useful methodology for addressing this omission.

Clearly, applying this methodology will lead to new questions about the relationship between tasks, performance and levels of proficiency. For example, what is the effect of task difficulty on performance? How do teachers respond to task difficulty? Do they take into account the difficulty of the task in assessing learner performance? In this paper we have attempted to provide a first step— a method of grading task difficulty in the CLB-- that may ultimately allow us to investigate these and other questions relating to our practice.

Conclusion

This paper has addressed issues arising from the use of the CLB in practice. By drawing on the relevant research literature from applied language studies and by examining the concept of task within the benchmark document, we have both problematized task progression and tasks difficulty and suggested potential approaches to address gaps in the CLB framework. It is important to note that these gaps which continue to challenge researchers, testers and teachers alike who are working with the CLB, are contentions and remain unresolved within the broader field of applied language studies itself.

References

Abedi, J. (2004). The No Child Left behind Act and English Language Learner: Assessment and Accountability Issues. *Educational Researcher,* 33(1), 4-14.

Abedi, J. Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content-based assessment: Analyses of extant data* (CSE. Tech. Rep. No. 603). Los Angeles: University of California, National Centre for Research on Evaluation, Standards, and Students Testing.

Adult Language Training Branch (2003). *Canadian Language Benchmarks* : *A summary of courses, programs, occupations and tests that have been benchmarked in Manitoba*. Winnipeg, MA: Government of Manitoba.

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing* 19(4), 453-476.

Bailey, C.(1998). *Learning about language assessment: Dilemmas, decisions, and directions.* Pacific Grove, CA: Heinle & Heinle.

Bygate, M, Skehan, P. and Swain, M. (Eds.). (2001). *Researching pedagogic tasks and second language learning, teaching and testing.* Essex, UK: Longman.

Breen, M. (1987). Contemporary paradigms in syllabus design. Parts 1 & 2. *Language Teaching*, 20(1), 81-92; (2), 157-174.

Brindley, G. (1989). *Assessing Achievment in the Learner-Centred Curriculum.* Sydney: National Centre for English Language Teaching and Research.

Brown, J.D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall.

Brown, J.D., Hudson, T., Norris, J. and Bonk, W. (2002). *An investigation of second language task-based performance assessments*. Honolulu, HI: University of Hawai'i Press.

Candlin, C. (2001). Afterword: Taking the curriculum to task. In *Researching pedagogic tasks: Second language learning, teaching and testing,* M. Bygate, P. Skehan and M. Swain, (Eds.), (pp. 229-243). Essex, UK: Longman.

Candlin, C. and Murphy, D. (1987). *Language Learning Tasks.* Englewood Cliffs, N.J.: Prentice-Hall International.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11(3), 367-383.

Cummins, J. (1984). *Bilingualism and Special Education: Issues in Assessment and Pedagogy.* Clevendon, England: Multicultural Matters.

Cummins, J. (1996), *Negotiating Identities: Education for Empowerment in a Diverse Society*. Toronto: California Association for Bilingual Education.

Ellis, R. (2003). *Task-Based Language Learning and Teaching.* Oxford: Oxford University Press.

Foster, P. & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition,* 18(3), 290-324.

Fox, J. (2004). 'Biasing for the best' in language testing and teaching: An interview with Merrill Swain. *Language Assessment Quarterly: An International Journal,* Vol. 1, No. 4, (pp. 235-251).

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, 13(1), 23-51.

Hajer, A. (2003) Using the Canadian Language Benchmarks: Implications for curriculum development and classroom practice. *Contact: Special Research Symposium Issue*, 30(2), 5-7.

Long, M. and Crookes, G.V. (1993). Units of analysis in syllabus design. In G. Crookes & S. Gall (Eds). *Tasks in a pedagogical context: Integrating theory and practice*. Clevedon: Multilingual Matters.

Long, M. and Robinson, P. (1998). Focus on form: Theory, research and practice. In C. Doughty and J. Williams (Eds.), *Focus on Form in Classroom SLA*. Cambridge: Cambridge University Press.

Loschky, L. and Bley-Vroman, R. (1993). Grammar and task-based methodology in S. Gass and E. Varonis (Eds.), *Talking to learn: Conversation in Second Language Acquisition.* Rowley, Mass.: Newbury House.

Messick. S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 23(2), 13-23.

Norris, J., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu, HI: University of Hawai'i Press.

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom.* Cambridge: Cambridge University Press.

Pawlikowska-Smith, G. (2002). *Canadian Language Benchmarks 2000.* Ottawa: Citizenship and Immigration Canada.

Prabhu, N.S. (1987). *Second Language Pedagogy*. Oxford: OUP.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning,* 45(1), 99-140.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics,* 22, 27-57.

Samuda, V. (2001). Guiding relationships between form and meaning during task performance: The role of the teacher. In M. Bygate, P. Skehan, and M. Swain (Eds.), Researching Pedagogic Tasks: Second language learning, teaching and testing, (pp. 119-140). Essex, UK: Longman.

Samuda, V. and Rounds, P.I. (1993). Critical episodes: Reference points for analyzing a task in action. In G. Crookes and S. Gass (Eds.), *Tasks in a Pedagogical Context: Integrating Theory and Practice*. Clevedon: Multilingual Matters.

Schmidt, D. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 17-46.

Short, D. (1993). Assessing integrated language and content instruction. TESOL Quarterly, 27(4), 627-656.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics,* 17, 38-62.

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics* 18, 268-86.

Skehan, P. (2003). Task-based instruction. *Linguistic and Language Teaching Abstracts,* 1-14.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook and B. Seidhofer (Eds.), *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press.

Swain, M. and Lapkin, S. (2001). Focus on form through collaborative dialogue: Exploring task effects. In M. Bygate, P. Skehan, and M. Swain (Eds.), *Researching Pedagogic Tasks: Second language learning, teaching and testing,* (pp. 99-118). Essex, UK: Longman.

Swales, J. (1987). *Episodes in ESP: A source and reference book on the development of English for science and technology*. Oxford, UK: Pergamon Press.

Witol, P. (2003) Incorporating competency based Canadian Language Benchmarks into a theme-based curriculum. *Contact: Special Research Symposium Issue*, 30(2), 15-22.

Yule, G. and Powers, M. (1994). Investigating the outcomes of task-based interaction. *System,* 22(1), 81-91.

Issues in Task-based Assessment of Communicative Competence

Gail Stewart

Introduction

Individuals whose professional lives are devoted to the delivery of ESL are held accountable by a broad range of stakeholders (Lacelle, Peterson & Rivera, 1994; Raimes, 1990; Shohamy, 1993), and crucial to this accountability is the development and implementation of valid and defensible instruments for determining placement, progress, outcomes and eligibility (Cumming, 1994; Darling-Hammond, 1994). Over the past decade, Canadian language assessment has begun to reflect an instructional evolution that has brought us from the darkness of the grammar translation period, through the fog of the audio-lingual method, and into the light of communicative methodology -- a transition that has led us to question our former reliance on objective, item-based measures and to strive toward more authentic, task-based models of assessment. This paper documents some of the complexities inherent in the development of task-based language assessment. An overview of the evolving context is followed by an examination of issues relating to authenticity, cultural accessibility, validity and reliability.

Shifting Paradigms and the Assessment Context

The concept of the learner as a central criterion in determining classroom objectives was advocated as early as the beginning of the 20th century, when ESL instruction was still floundering in the grammar translation approach. Dewey (1916) stressed the importance of democracy within the educational system and asserted that a crucial aspect of that democracy was a thoughtful consideration of learner needs and goals. In a democracy, it is important to give newcomers the tools that empower them to function successfully in the new society. One of the most important of these tools is language. Because relationships of power are closely associated with language usage (Rockhill & Tomic, 1994), it is essential for the learner to master the dominant language. Without proficiency in the language of the new culture, the learner is unable to take part in the basic day-to-day tasks and routines that constitute a normal, productive life.

The role of culture in preparing learners to participate effectively in society should not be underestimated.

> Social efficiency as an educational purpose should mean cultivation of power to join freely in shared or common activities. This is impossible without culture, while it brings a reward in culture, because one cannot share in intercourse with others without learning -- without getting a broader point of view and perceiving things of which one would otherwise be ignorant. And there is perhaps no better definition of culture than that it is the capacity for constantly expanding the range and accuracy of one's perception of meanings. (Dewey, 1916, p. 123)

The association which Dewey makes between culture and one's "perception of meanings" is fundamental to second-language education. An understanding of cultural nuance is essential because native speakers attend and respond not merely to words but to the illocutionary force of a message. Simple words have various meanings and implications depending upon the context in which they are used. Thus, learners not only must master the linguistic features of discourse but also be made aware of the "rules of illocutionary and social competence" (Richards, 1988, p. 86). It, therefore, becomes apparent that, alone, the assimilation of grammatical and structural rules and the rote repetition of stock phrases cannot equip a newcomer to deal with the complexity of functioning in a new environment. Under the grammar translation and audio-lingual paradigms, the subtle but essential paralinguistic and contextual aspects of language remained a mystery.

By the early 1960's, second-language educators had begun to realize that classroom instruction was not preparing adult learners to function optimally in their newly-chosen lives. As a result, that decade saw the audio-lingual approach supplanted by the communicative movement, in which previous models of general language proficiency were finally replaced by a philosophy that now acknowledges the centrality of the learner, the importance of culture, and the need for a balanced focus in each of four skill areas -- reading, writing, listening, and speaking. Under the communicative paradigm, the approach to teaching methodology is learner-centred, and the primary focus is to address those aspects of language that learners require to negotiate the intricacies of day-to-day living in a new culture. Language is no longer stripped of its context, but rather has come to be viewed as inextricably linked to the culture in which it has developed. As Valette (1971) notes:

> the new language is not simply a "code", ... but an integral part of a culture ... Now, a balance is being sought between the spoken language and the written language, culture and civilization ..., cognitive processes and the acquisition of habits. (pp. 816-819)

This communicative approach to second-language instruction brings the ESL field into alignment with Dewey's (1940) notion of language as a vehicle for social interaction:

> It is true that language is a logical instrument, but it is fundamentally and primarily a social instrument. Language is the device of communication. (p.11)

Defining the Communicative Construct

With the emergence of a communicative paradigm, the view of language as a unitary construct has been rapidly succeeded by multidimensional models that take into account the functional and contextual aspects of communication. Canale and Swain (1980) and Swain (1984) posit a four-dimensional model comprising linguistic, discourse, strategic and sociolinguistic competencies, while Bachman (1988) presents a three-pronged approach that includes language competence, strategic competence, and

psychophysiological mechanisms. In relating these newer models to language assessment, Wesche (1983) states:

> In this perspective, language competence is viewed as a complex system of rule sets which operate simultaneously at many levels to determine the organization of grammatical forms for the fulfillment of communicative and other language functions ... Language testing which does not take into account propositional and illocutionary development beyond the sentence level as well as the interaction between language behaviour (verbal and non-verbal) and real-world phenomena is, at best, getting at only a part of communicative competence. (p. 42)

The range of behaviour suggested by the new models of communicative competence is captured in the Canadian Language Benchmarks 2000 (Centre for Canadian Language Benchmarks, 2000). Describing a hierarchy of skills and strategies applied within a sociocultural context for the accomplishment of specific essential tasks, the document is intended to reflect the increasing linguistic demands placed on Canadian immigrants as they attempt to settle and integrate in the new culture. Development of the Canadian Language Benchmarks (CLB) has been undertaken in the spirit of transparency and accountability outlined by Russell (1993):

> Benchmarks and standards focus on curriculum substance in an open manner that permits and, perhaps, encourages public debate about what students are taught and what they learn, as well as what general directions education should take. (p. 1)

These Canadian standards serve three important purposes. First, they provide a common language for the discussion of learner needs and objectives. Second, they are intended to serve as a systematic framework for understanding the progression of communicative competence and for planning ESL programs and curricula. Third, they are expected to provide a clear definition of the communicative construct and a description of the corresponding domain that can be applied in the development of assessment instruments.

Issues of Authenticity and Accessibility

The shift toward a communicative instructional methodology has necessarily brought about a reconsideration of approaches to ESL assessment. The communicative domain comprises a range of behaviours associated with contextualized performance on tasks requiring a practical demonstration of competence. It, therefore, follows that the most desirable forms of assessment would be those that demonstrate a high degree of authenticity by "tackling the issues at hand" (Shelley & Cohen, 1986, p. 82). At a fundamental level, the issue of authenticity is tied directly to face validity, as there is no doubt that assessments boasting a high degree of apparent relevance are more likely to be widely accepted and approved by the general public. Arguably, face validity is a desirable feature in that it can encourage stakeholder buy-in and increase the likelihood that examinees will cooperate in an assessment procedure. However, it is not technically a form of validity that contributes in a meaningful way to the soundness or reliability of

an assessment; nor is it identified as a requirement in the commonly accepted standards for educational measurement (Joint Advisory Committee, 1993; Joint Committee on Testing Practices, 1988).

Within the ESL profession, there is a noted tendency to assume that test development procedures which are geared toward creating authentic content and replicating real-life circumstances produce results that are valid by definition (McNamara, 1995a), even though the very notion that a language test can be authentic is open to debate. Achieving authenticity is more problematic than one might imagine. It is difficult in the context of an assessment to take into account, for example, certain aspects of the criterion setting which may be important, such as physical distance and ambient noise (c.f. Bailey, 1985). Furthermore, role-plays or simulations carried out under assessment conditions cannot possibly target the layers of complexity that characterize real-life relationships between, for example, a doctor and a patient or a teacher and a student. Spolsky (1985) suggests that

> any language test is by its very nature inauthentic, abnormal language behaviour. With examinees who do not know the rules of the game or who are unwilling to play according to them, the results will not be an accurate and valid account of their knowledge. (p. 39)

Wiggins (1993) claims that this is not necessarily the case. Though he admits that a test is indeed "a contrivance", he argues that

> it should not feel like one or be perceived as one ... A context is thus realistic to the extent that we so accept the premises, constraints, and "feel" of the challenge that our desire to master it makes us lose sight of any extrinsic or contrived factors at stake -- factors such as the reality that someone is evaluating. (p. 232)

This is an appealing premise, but a tall order for test development. The notion that a communicative procedure, administered under test conditions, could potentially become so engaging and absorbing that an examinee would very nearly forget about the reality of the circumstances, is probably not realistic. Test developers who are inclined to hope otherwise would do well to remember the words of Lantolf and Frawley (1988), who state in reference to speaking assessment, "in essence, there is only one task in oral proficiency -- the test" (p. 183).

To the extent that it is possible to create assessment procedures that accurately replicate real situations and circumstances, issues arise with regard to fairness and equity. When an assessment does not favour all groups equitably, it is sometimes referred to as unfair or biased. In measurement, a distinction is often made between the two terms. Whereas 'bias' applies to a quality of the instrument being used, 'unfairness' pertains to the purpose to which the test is assigned (Shepard, 1982). Cultural bias can be defined as an aspect of the test result that "may feed back information about your cultural background

instead of your ability" (Strenio, 1981, p. 201). Assessment that is biased or unfair will have limited validity. However, the attempt to create bias-free assessments may not be compatible with a task-based model of communicative competence. Because it is next to impossible to elicit language performance for the accomplishment of real tasks without introducing some degree of bias, a fundamental paradox arises. What is perceived to be the greatest strength of the task-based approach is also its greatest weakness (Peirce & Stewart, 1997).

This paradox is particularly evident in the controversy that surrounds task-based performance assessment. Jones (1985) defines a performance assessment as one that

> measures performance on tasks requiring the application of learning in an actual or simulated setting. Either the test stimulus, the desired response, or both, are intended to lend a high degree of realism to the test situation. (p. 16)

One of the main difficulties in task-based assessment is determining how to deal with the unavoidable expectation that examinees should have knowledge pertaining to situation, context and culture. While the field is deeply concerned with issues of equity, we also acknowledge that socio-cultural awareness is, to some degree, a necessary component of high-level communicative performance (Loughrin-Sacco, 1990). One possible solution is to strive for tasks that are "equitably accessible" (Peirce & Stewart, 1997) rather than culture-free. To be deemed equitably accessible, tasks undergo a rigorous process of review by a cultural advisory committee representing the range of groups to be assessed. This committee ensures that, while certain task types may be unfamiliar to some examinees, the assessment content is structured such that a learner is able to access meaning and respond in a manner commensurate with his or her level of competence in the target skill.

Issues Relating to Validity

The communicative backlash against standardized, indirect measures of proficiency has propelled a movement toward tests that "appeal to educators who see the need to 'get real'" (Frechtling, 1991, p. 24). However, concerns raised in the previous section underscore the insufficiency of an approach that presumes validity based on the fact that an instrument demonstrates a high degree of authenticity. Caught in the throes of shifting paradigms, we must take care to ensure that our desire to create something new, authentic, and ostensibly 'better' does not supersede the need to address important issues that pertain to validity. As Ochsner (1979) once observed, in reference to language researchers, "If chemists juggled their basic units like we do, their laboratories would blow up" (p. 58). Alderson (1986) agrees that constant change is not always a good thing and suggests that, before the latest innovation is adopted, there should be solid psychometric evidence to prove that it can do the job better than the methodology it is intended to replace.

The reluctance in the field to accept the validity of indirect measures is based on concerns that such indicators are limited in the extent to which they can predict such facets of the construct as fluency and interactive ability. Wiggins (1993) summarizes the perceived drawbacks of standardized indirect assessment thus:

> the forms of testing and scoring used are indirect and generic, designed to minimize the ambiguity of tasks and answers ... But such forms of testing simply do not tell us what we need to know -- namely, whether students have the capacity to use wisely the knowledge they have. This is a judgement that we can make only through tasks that require students to "perform" in highly contextualized situations. (p. 208)

Bachman and Savignon (1986), however, caution that the quality of inferences based on performance in context should not be overestimated. These authors are concerned that the use of the term 'direct' can be misleading. It is often assumed in a direct assessment that "the surface expression of competence" (Wood, 1981, p. 139), or the observed indicator, is, in fact, the construct. Bachman and Savignon point out that this is not the case -- that the construct itself is never actually observed in an assessment, regardless of the method or the conditions of administration. They state:

> As with all mental measures, language tests are indirect indicators of the underlying traits in which we are interested. That is, we are generally not interested so much in how individuals perform on a given test on a given day, as we are in their ability to use language communicatively at different times and in a wide range of contexts.
>
> (p. 382)

Because of the consequences for stakeholders resulting from decisions made on the basis of language assessment, the concept of validation is "ominously important" (Cumming, 1996, p. 1). However, it is not unusual to find conflicting views as to what constitutes an appropriate definition of the term 'validity'. In many cases, the notion of multiple validities documented historically in the literature (Angoff, 1988) still prevails, leading to confusion about the relative importance of one 'type' of validity over another and to erroneous notions about the significance of face validity, authenticity and wash-back. Even in cases in which a unified definition of validity (Cronbach, 1980, p. 99) is adopted, a task-based approach complicates test development because authentic tasks have a tendency to take centre stage, sometimes overshadowing the construct to the point in which it seems as though content is all that matters. Messick (1989) states that content validity lies, not within the test itself, but in "the relationship between the test and the domain of reference" (p. 41) and that "there is often no sharp distinction between test content and test construct" (p. 36). Pedhazur and Schmelkin (1991) claim that "content validity is not a type of validity at all" (p. 79), while Shepard (1993) acknowledges the importance of content but agrees with Messick that it should not be the sole basis for establishing validity.

It may matter little whether or not we choose to adhere to a philosophy of multiple validities if we determine to keep in mind that the heart of validation is the interpretation of test scores for decision-making purposes. Every task-based assessment should be designed with its specified purpose in mind and should be validated strictly in accordance with that intended objective. In other words, the simple fact that an instrument has been developed with integrity and analyzed soundly does not make it inherently valid. Wood (1987) states that "a test cannot be valid in general; it is valid for a purpose" (p. 136). Moreover, it is not the test per se that is said to be valid, but rather the interpretation of data which are gathered by means of the assessment procedures. Cronbach (1971) stresses:

> Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test is valid. (p. 447)

The interpretation of test results demands a careful synthesis of indicators. If the objective is to determine whether a candidate can satisfy eligibility requirements based on some specified criterion, then the standard-setting procedures employed to derive reasonable cut-off scores need to be examined as part of the validation process. In addition, 'decision consistency' (Traub, 1994; Traub & Rowley, 1980), the degree to which the results of repeated testing would render the same ultimate decision for a given examinee, has to be investigated.

While the responsibility for standard setting rests in the hands of the test user rather than the test developer, the user's tendency to look for quick solutions to important decision-making problems should be kept in mind when considering validity. Funding bodies often have high expectations for the potential application of test instruments and little understanding of the intricacies of the validation process. Language test validation requires a complex integration of many types of evidence. As Cumming (1996) indicates:

> establishing validity in language assessment is by all accounts problematic, conceptually challenging, and difficult to achieve -- probably more so than is recognized outside the specialized spheres of those few persons who make this endeavor their business. (p. 1)

Validity in testing sits at the core of accountability to the ESL learner, to the field and to society. As language assessments are increasingly used for high-stakes decisions that affect eligibility for employment, university entrance and even immigration, it is essential that greater emphasis be placed on evaluating the appropriateness of any instrument for its intended application. The validation process should expose the limitations as well as confirm the strengths of a given assessment procedure, and, in consideration of all relevant factors that might affect accountability, bring test users and test takers to a full awareness of the types of conclusions and decisions that the interpretation of test results can reasonably and defensibly support.

Issues Relating to Reliability

In moving beyond a previous reliance on indirect measures toward assessments that are more immediate and authentic, we have to acknowledge that, to some degree, reliability is being sacrificed. At the same time, there seems to be a sense of acceptance that, despite this one unfortunate limitation, finally, tests are becoming more valid, and validity is, after all, far more important than reliability. Unfortunately, the tendency to assume that validity can exist in the absence of reliability is based on a fallacy. Reliability is an indispensable component in validity. Though there are some who would position reliability as "an option rather than a requirement" (Moss, 1994), it is generally agreed among measurement professionals that there can be no valid test that is not also reliable (Henning, 1987: Pedhazur & Schmelkin, 1991; Popham, 1978).

In task-based assessment, the test developer attempts to situate the expression of competence in such a way as to offer the learner an opportunity to deal with the "real, 'messy' uses of knowledge in context -- the 'doing' of a subject" (Wiggins, 1993, p. 207). In so doing, the practitioner encounters an inescapable tension between these objectives and the requirement for reliably measurable outcomes. While the development of any language assessment is a struggle in complexity and contradiction (Wesche, 1987), the necessity to apply a task-based approach further complicates the undertaking (Mislevy, Steinberg & Russell, 2002). Even in cases in which the stakes are relatively low and decisions easily reversible, it is important that an assessment instrument be psychometrically sound and valid for its intended purpose. In situations where examinees' futures hang in the balance or public safety is a concern, the imperative to ensure the reliability of an instrument is compounded.

The term 'reliability' refers to consistency of measurement. According to classical test theory, any score attained on a test instrument is composed of a true score element and some degree of random error. A psychometric estimate of the size of the error for any given instrument provides an indication of that instrument's reliability. Generally speaking, the smaller the error, the more reliable the instrument (Traub, 1994). The traditional definition of reliability requires that an assessment procedure make distinctions among examinees with respect to their placement on a continuum of skill or competence (Traub & Rowley, 1980), but devising a task-based continuum that accurately follows the progression of developing language competence is a challenging undertaking. The approach taken in the CLB is to situate descriptors of construct-related behaviour within the context of performance on authentic tasks considered to be typically suited for each benchmark level but not empirically exclusive to any one benchmark.

In working with the CLB, it is often difficult to separate behaviours from tasks, but it is essential to do so because a single task can never reliably be pegged to a single benchmark. While a task such as, "Give complex instructions on familiar first aid and emergency procedures in the work place" (Centre for Canadian Language Benchmarks, 2000, p. 116), located at speaking benchmark 9, may typify the sort of content that is appropriate for learners at that level, it would be difficult to justify in all cases that this task is more difficult than, "Explain how to avoid jet lag, stay on a budget, or handle a household emergency" (Centre for Canadian Language Benchmarks, 2000, p. 66),

located at benchmark 7. From an assessment perspective, then, tasks are best viewed not as exclusive indicators of benchmark status, but as springboards for eliciting a range of observable performance. In the interest of fairness to examinees, the broadest possible variety of representative task types and genres should be presented and evaluated.

In terms of statistical analysis, methodologies based on classical reliability theory can be applied when the approach to testing is indirect and the items objective. In other cases, it is necessary to apply alternate statistical methods, which take into account the many variables that can often influence task-based assessment. Clemens (1971) identifies administrative procedures as a source of variability that can potentially threaten reliability:

> Every aspect of the setting in which the test is given and every detail of the procedures may have an influence on performance and hence on what is measured (p. 449)

Task-based assessment can prove particularly problematic in this regard. For example, in many cases, the examinee responds to a live interlocutor, whose elicitation techniques may vary from one administration to another. If the interaction takes the form of a conversation, the examinee is usually expected to take unnaturally long turns so that administration time is not wasted. Attempts to standardize administration time can work against the facilitation of optimum performance, as it takes some candidates longer than others to demonstrate the full extent of their proficiency. If examinees are expected to interact in some manner with one another, elements of one examinee's performance are likely to be dependent to some degree on the strategies and abilities of one or more of the other candidates. In a role-play or simulation, the conditions of administration and the quality of the interaction are likely to exhibit characteristics of the criterion situation and behaviour that are difficult to control. In fact, generally speaking, the more closely an assessment approximates the characteristics of a real-life situation, the less reliable the measurement becomes (Fitzpatrick & Morrison, 1971). As Wiggins (1993) cautions

> we should keep the test maker's dilemma in mind: fidelity to the criterion situation maximizes the complexity and ambiguity of the task requirements and the freedom to respond as one sees fit -- conditions that work against maximizing standardization and reliability. (p. 208)

Subjective scoring is susceptible to random error which may be introduced by unpredictable rater idiosyncrasies (Henning, 1987). If an acceptable degree of reliability is to be achieved, assessors must be trained to rate with consistency. Traub (1994) states:

> To the extent that the judgements of one scorer vary unsystematically from those of another scorer, or to the extent that the judgements of one scorer vary unsystematically over time, then the score assigned a particular response will depend on the particular individual assigned the task of scoring a response or,

> in the event that only one judge is involved, and scoring extends over a period of time, the score will depend on the judgmental frame of reference the judge brings to bear on the response when the scoring is done. (p. 109)

This statement underscores the two crucial components necessary for consistency in subjective rating. One component is intra-rater reliability -- the degree to which a single rater assigns the same score to a given candidate over two or more administrations. The other component is inter-rater reliability -- the degree to which two or more raters assign the same score to a given candidate. Ensuring either type of reliability is dependent on a number of requirements. The first is the quality of the scoring bands or rating scales. The descriptors need to be sufficiently unambiguous for assessors to make judgements with confidence and consistency. This is difficult to achieve because the observations made in a language assessment comprise multiple indicators. Often it is not possible to take into account all of the ways in which the combined indicators could reasonably be interpreted. In addition, it is difficult to determine how much emphasis, if any, should be placed on non-linguistic criteria (Brown, 1995).

An analytic component is sometimes embedded in a holistic scale as a way of attempting to disambiguate it (Douglas & Selinker, 1992). A general statement describes the most salient features of the band or level, and then a more detailed explanation is given of the observational components that apply. Sometimes, there is a series of scales -- one for each proposed facet of the construct. In such a case, consideration must be given to the manner in which final results are to be reported. If language proficiency is indeed a multidimensional construct, it may be counterproductive to present a single score as an indicator without some reference to an examinee's performance in the various component areas (Masters, 1990). It may be equally problematic to simply provide ratings on the components without also offering information on how these ratings should be synthesized and interpreted (de Jong, 1990). Because of this difficulty with rating scales, some test developers present assessment results in the form of checklists, which may be helpful for diagnostic purposes, but are not particularly useful if the objective is to rank candidates.

Because it is difficult to write scoring bands that can be interpreted and applied in only one way, the second requirement for ensuring reliability is rigorous training of the assessors, or raters, who will use the procedures. A training session in which assessors are introduced to the scoring procedures and then required to practise rating samples before going into the field can help to ensure that all raters apply the scales or bands in an accurate and consistent manner (Lumley & McNamara, 1995). During training, every effort is made to guide raters in the systematic application of criteria, so that, ultimately, all assign the same weighting to the linguistic and non-linguistic features of performance. Unfortunately, this ideal is not always achieved. Evidence suggests, for example, that raters are influenced by elements of discourse that are not included in rating scales (Douglas, 1994). Furthermore, studies show that raters who are native speakers of English do not employ scales in the same way that non-native speakers do and that ESL professionals rate differently than assessors from other occupational backgrounds (Brown, 1995). The implication is that two different raters may make the same

judgement, but for different reasons. Very often then, even when consistency is achieved, accuracy is not.

A further consideration in the reliability of task-based assessment is test length. Concern exists around the limited range of linguistic behaviour that can be reasonably elicited and observed during the time allotted for communicative assessment (Mehrens, 1992; Perrett, 1990). Since, in general, reliability is improved as a result of an increase in the number of test items (Traub & Rowley, 1991), there is a greater likelihood that an acceptable degree of reliability will be achieved in an objective test in which more territory can be covered in a shorter length of time. Establishing reliability in an objective test usually involves an examination of the correlations among items. Whereas developers of objective tests run the risk of discarding items that correlate poorly with the pool, thereby creating tests that are less representative of the domain (Cronbach, 1971), developers of task-based instruments may incorporate so many task-related requirements that the resulting instrument exceeds the parameters of the original domain without, in fact, adequately addressing all of the elements implied by the domain definition.

In terms of reliability, the trend toward task-based assessment introduces many potential sources of error, a complexity that has led to an extension of classical reliability theory referred to as generalizability theory, or G theory. The key consideration in G theory is the extent to which observed indicators on a given occasion in a particular setting allow us to make generalizations about behaviour under other circumstances (Shavelson & Webb, 1991). G theory is based on the assumption that an examinee's performance is a function of various factors or facets that exist independent of the construct (Bachman, Lynch & Mason, 1995). These facets might include such elements as the task, the test method, the rater and certain aspects of the administrative conditions. Because scores fluctuate as a result of the influence of these facets, generalizability theory enables researchers to

> estimate what the true score or universe score would have been over an infinite number of observations under various conditions of different factors or facets. (Bolus, Hinofotis & Bailey, 1982, 248)

Since "validity is more important than reliability, but is impossible without it" (Wood, 1987, p. 136), an examination of evidence of reliability should be the first step in evaluating the quality of an assessment instrument. According to Walsh and Betz (1990), reliability places restrictions on the potential validity of any test, establishing an upper limit on the degree of validity that is possible. The level of reliability that is acceptable varies from one instrument to another, depending upon the intended purpose. Assessments can range from "quick and dirty", wherein the consequences of misjudgement are negligible, to "the full bore", which can have serious effects on the lives of examinees (McNamara, 1995b.). Tinkelman (1971) states:

The minimum acceptable reliability depends on the seriousness of the decisions

> to be made about examinees. If the purpose of the test is to place students in broad instructional groups, with ample opportunity for shifting from one group to another as learning progresses, the consequences of faulty measurement may not be critical. If the cutoff score is to be used to award jobs or to grant scholarships, reliable measurement may be critical indeed. (p. 71)

Conclusion

The communicative movement has brought about an emphasis on the use of language in context for the accomplishment of authentic tasks, and as a result, ESL test development has experienced a metamorphosis. Traditional, indirect assessment methods have largely been replaced by alternatives based on linguistic models of communicative competence. The goals associated with communicative assessment are noble, born of an educational movement that places learner needs at the core of the curriculum. This approach is healthy for individual learners and society.

A democratic approach to assessment demands attention to the consequences that testing practices may have for all stakeholders, and, as language tests are increasingly used for high stakes decisions, it is essential that emphasis be placed on evaluating the appropriateness of any instrument for its intended application. Developers of communicative tests should, therefore, give serious consideration to the concerns of the measurement community by striving to design instruments that are as valid and reliable as possible while retaining the authenticity and relevance that are central to the communicative approach. In turn, the measurement community should broaden its perspective to include a consideration of alternate forms of interpretation that can address the unique requirements of task-based assessment.

The endeavour to create an infrastructure that supports the ongoing development and validation of task-based assessment is well worth the time and the effort. As Eisner (1985) notes:

> Evaluation as a process can perform many different functions in education: it can reward, it can screen, it can select, it can at times help us determine if our objectives have been achieved. But perhaps the most important function of evaluation from a strictly educational point of view is to help educators improve the quality of educational life for students. There are no short-cuts that I know of to this end. (p. 187)

An educational system that places learner needs at its centre does not seek out short-cut solutions to long-term objectives. Task-based assessment may not be as neat and uncomplicated as the methods that have preceded it, but neither is language learning a straight-forward process. If the ESL field is to be accountable to learners and to the community, our assessment procedures must address the real objectives that ESL learners

face in preparing to negotiate the complexity and diversity of authentic communication. It is thus the responsibility of those who support democracy in education to continue the pursuit of a framework, a terminology and a research methodology to establish fully and defend the legitimacy of task-based assessment.

References

Alderson, J.C. (1986). Innovations in language testing? In M. Portal (Ed.), *Innovations in Language Testing*. (pp. 93-108). Windsor: NFER-Nelson.

Angoff, W. (1988). Validity: an evolving concept. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 19-32). Hillsdale: Erlbaum.

Bachman, L.F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition*, 10,(2), 149-164.

Bachman, L.F., Lynch, B.K. & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12,(2), 238-257.

Bachman, L.F. & Savignon, S.J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *Modern Language Journal,* 70,(4), 380-390.

Bailey, K.M. (1985). If I had known then what I know now: performance testing of foreign teaching assistants. In P.C. Hauptman, R. LeBlanc & M.B. Wesche (Eds.), *Second Language Performance Testing* (pp. 153-180). Ottawa: Ottawa University Press.

Bolus, R.E., Hinofotis, F.B. & Bailey, K.M. (1982). An introduction to generalizability theory in second language research. *Language Learning,* 32,(2), 245-258.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12, 1-15.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1,(1), 1-47.

Centre for Canadian Language Benchmarks (2000). *Canadian Language Benchmarks: English as a Second Language for Adults and English.* Ottawa: Citizenship and Immigration Canada.

Clemens, W.V. (1971). Test administration. In R.L Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 188-201). Washington, DC: The American Council on Education.

Cronbach, L.J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement,* 5, 99-108.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 443-507). Washington, D.C.: American Council on Education.

Cumming, A. (1996). Introduction: The concept of validation in language testing. In A. Cumming. & R. Berwick (Eds.), *Validation in Language Testing* (pp. 1-15). Clevedon: Multilingual Matters.

Cumming, A. (1994). Does language assessment facilitate recent immigrants' participation in Canadian society? *TESL Canada Journal*. 2,(2), 117-133.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review,* 64,(1), 5-29.

de Jong, J.H.A.L. (1990). Response to Masters: linguistic theory and psychometric models. In J.H.A.L. de Jong & D.K. Stevenson (Eds.), *Individualizing the Assessment of Language Abilities.* Clevedon: Multilingual Matters.

Dewey, J. (1940). *Education Today*. New York: Greenwood Press.

Dewey, J. (1916). *Democracy and Education: An Introduction to the Philosophy of Education.* New York: The Free Press.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, 11,(2), 125-144.

Douglas, D. & Selinker, L. (1992) Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, 20,(3), 317-328.

Eisner, E.W. (1985). *The Art of Educational Evaluation: A Personal View*. London: The Falmer Press.

Fitzpatrick, R. & Morrison, E.J. (1971). Performance and product evaluation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 237-270). Washington, DC: American Council on Education.

Frechtling, J.A. (1991). Performance assessment: moonstruck or the real thing? *Educational Measurement: Issues and Practice*, 10,(4), 23-25.

Henning, G. (1987). *A Guide to Language Testing*. Rowley: Newbury House.

Hopkins, K.D., Stanley, J.C. & Hopkins, B.R. (1990). *Educational and Psychological Measurement and Evaluation* (7th ed.). Englewood Cliffs: Prentice-Hall.

Joint Advisory Committee. (1993). *Principles for Fair Student Assessment Practices for Education in Canada*. Edmonton: Centre for Research in Applied Measurement and Evaluation.

Joint Committee on Testing Practices. (1988). *Code of Fair Testing Practices in Education*. Washington, D.C.: American Psychological Association.

Jones, R.L. (1985). Second language performance testing: an overview. In P.C. Hauptman, R. LeBlanc & M.B. Wesche (Eds.), *Second Language Performance Testing* (pp. 15-24). Ottawa: University of Ottawa Press.

Lacelle-Peterson, M. & Rivera, C. (1994). Is it real for kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review*, 64, 55- 75.

Lantolf, J.P. & Frawley, W. (1988). Proficiency: understanding the construct. *Studies in Second Language Acquisition,* 10, 181-195.

Loughrin-Sacco, S.J. (1990). The reflections and confessions of an ACTFL Oral Proficiency Interview workshop participant. *The Canadian Modern Language Review,* 46,(4), 706-713.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12,(1), 54-71.

Masters, G.N. (1990). Psychometric aspects of individual assessment. In J.H.A.L. de Jong. & D.K. Stevenson (Eds.), Individualizing *the Assessment of Language Abilities* (pp. 56-71). Clevedon: Multilingual Matters.

McNamara, T.F. (1995a.). Modeling performance: opening Pandora's box. *Applied Linguistics,* 16,(2), 159-179.

McNamara, T.F. (1995b.). Research and practice in testing oral skills. Plenary address at the International Colloquium on Second Language Speaking and Writing: Research - Teaching - Technology, University of Ottawa.

Mehrens, W.A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, 11,(1), 3-9.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: MacMillan.

Mislevy, R.J., Steinberg, L.S. & Russell, G.A. (2002). Design and analysis in task-based language assessment. *CSE Technical Report 579*. Los Angeles: University of California.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, 23, (2), 5-12.

Ochsner, R. (1979). A poetics of second-language acquisition. *Language Learning*, 29, 53-80.

Pedhazur, E.J. & Schmelkin, LP. (1991) *Measurement Design and Analysis: An Integrated Approach*. Hillsdale: Lawrence Eilbaum.

Peirce, B.N. & Stewart, G. (1997). The development of the Canadian Language Benchmarks Assessment. *TESL Canada Journal,* 14,(2), 17-31.

Perrett, G. (1990). The language testing interview. In J. de Jong, & D.K. Stevenson (Eds.), *Individualizing the Assessment of Language Abilities*. Clevedon: Multilingual Matters.

Popham, J. (1978). *Criterion-Referenced Measurement*. Englewood Cliffs: Prentice-Hall.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly*, 24, 427-442.

Richards, J.C. (1988). Talking across cultures. In G.H Irons (Ed.*), Second Language Acquisition: Selected Readings in Theory and Practice* (pp. 84-94). Welland: Canadian Modern Language Review.

Rockhill, K. & Tomic, P. (1994). Situating ESL between speech and silence. In J. Willinsky & J. Gaskell (Eds.), *Gender In / Forms Curriculum: From Enrichment to Transformation*. New York: Teachers College Press.

Russell, H. (1993). Standards, Benchmarks, and Ontario Curriculum: Introduction. *Orbit*, 24,(2), 1.

Schmidt, R. (1983). Interaction, acculturation, and the acquisition of communicative competence: A case study of one adult. In N. Wolfson, & E. Rowley, MA: Newbury House.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury Park: Sage.

Shelley, D. & Cohen, D. (1986). *Testing Psychological Tests.* New York: St. Martin's Press.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education,* 19, 405-448.

Shepard, L.A. (1982). Definition of bias. In R.A. Berk. (Ed.), *Handbook of Methods for Detecting Test Bias* (pp. 9-32). Baltimore: Johns Hopkins University Press.

Shohamy, E. (1993). The power of tests: the impact of language tests on teaching and learning. *NFLC Occasional Papers*, Washington, D. C.

Spolsy, B. (1985). The limits of authenticity in language testing. *Language Testing*, 2,(1), 31-40.

Strenio, A.J. (1981). *The Testing Trap.* New York: Rawson, Wade Publishing Inc.

Swain, M. (1998). The output hypothesis, second language learning and immersion education. In J. Arnau & J. Artigal (Eds.), *Immersion programs: A European perspective* (pp. 127-140). Barcelona, Spain: Publicacions de la Universitat de Barcelona.

Swain, M. (1984). Teaching and Testing Communicatively. *TESL Talk,* 15, (1&2).

Tinkelman, S.N. (1971). Planning the objective test. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 46-80). Washington, D.C.: American Council on Education.

Traub, R.E. (1994). *Reliability for the Social Sciences: Theory and Applications*. Newbury Park: Sage.

Traub, R.E. & Rowley, G.L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, 10,(1), 37-45.

Traub, R.E. & Rowley, G.L. (1980). Reliability for test scores and decisions. *Applied Psychological Measurement*. 4,(4), 517-545.

Valette, R.M. (1971). Evaluation of learning in a second language. In B.S. Bloom, J.T. Hastings & G.F. Madaus (Eds.), *Handbook on Formative and Summative Evaluation in Student Learning* (pp 815-508). Toronto: McGraw-Hill.

Walsh, W.B. & Betz, N.E. (1990). *Tests and Assessment* (2nd ed.). Englewood Cliffs: Prentice-Hall.

Wesche, M.B. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing*, 4,(1), 28-47.

Wesche, M.B. (1983). Communicative testing in a second language. *Modern Language Journal*, 67,(1), 41-55.

Wiggins, G.P. (1993). *Assessing Student Performance*. San Francisco: Jossey-Bass Inc.

Wood, R. (Ed.), (1987). *Measurement and Assessment in Education and Psychology*.

Reflecting on Classroom Communication

Thomas S. C. Farrell

Introduction

In recent times teachers have been encouraged to reflect upon every aspect of their teaching (Farrell, 2004a) because as Richards and Lockhart (1994) suggest, teachers are often unaware of what they do when they teach and how their teaching influences their students' levels of learning. However, one aspect of reflection, which surprisingly has been neglected in the literature is classroom communication. Thus, teachers are often unaware of what communication patterns exist in their classes. For example, they may not be aware of the various ways classroom communication opens or blocks opportunities for students to reach optimum levels of learning (Farrell, 2004b). Additionally, language teachers may not even know how to investigate various communication patterns in their classrooms. The purpose of this paper is to explore briefly how language teachers can reflect on some different classroom communication patterns that exist in their classrooms.

The Nature of Classroom Communication

Classroom communication, that is, face-to-face communication and interaction between teachers and students, is shaped by moment-to-moment actions and interactions within that classroom (Johnson, 1995). This communication may seem chaotic to outsiders, but, in fact, is ordered, and, it is the teachers, because of their status and power, who are mostly responsible for setting up the type of communication in their classrooms, which, in turn, has enormous influence upon their students' levels of learning. For example, the teacher decides the topic that will be discussed, as well as who will participate and when. How a teacher chooses, (often subconsciously), to organize this communication depends on, and reflects his/her prior experiences as a student of language or other), as well as his/her theories and beliefs about language learning and teaching. Consequently, there is an important need for language teachers to be able to recognize and understand how these communication patterns influence (positively or negatively) their students' learning.

That said, it is equally (and some may even say *more* important) for students to recognize patterns of classroom communications established by the teacher so that they will become more aware of teacher expectations, especially in terms of their role in the classroom (Farrell, 2004b). For example, students must be able to identify what, why, when and how they are expected to communicate in the classroom (Mehan, 1979) in order to be communicatively competent in the classroom (Johnson, 1995). As Barnes (1976, p. 33) points out, classroom communication as a system "is a matter not only of how the teacher sets up classroom relationships and discourse, but also of how the pupils interpret what the teacher does." Barnes (1976) further suggests that the success students achieve in a classroom depends upon various patterns of communication that they are able to recognize and that are set up by the teacher. To create a successful learning environment, both teachers and students must take responsibility for recognizing various structures underlying communication in their classrooms.

Underlying Structures of Classroom Communication

Classroom communication has been called a "game" (Belleck et. al., 1966) in which the teacher sets up rules and communication that tend to follow established patterns after the teacher has gained valuable teaching experience. The teacher, either consciously or unconsciously (the latter is most prevalent), uses language to establish these patterns to perform two functions simultaneously: to carry the message that a teacher wants to communicate and, at the same time, to convey specific information about who the teacher is, and whom he/she is talking to. The following pattern is the unmarked (usual) pattern of classroom communication that is found in many language classrooms (Farrell, 2004b): the teacher initiates (I), a student (or students) responds (R), and this is usually followed by the teacher's evaluation (E) to the response, or simply the *IRE*. The following example (Episode 1) illustrates this unmarked underlying communication structure that can be found in a majority of language classrooms:

Episode 1: Unmarked Classroom Communication Structure

1: Teacher: What day is today?	[**I**nitiation]
2: Students: Monday.	[**R**esponse]
3: Teacher: Very good. Monday.	[**E**valuation]

In turn 1, the teacher asks the students what day it is during a lesson on the days of the week. The students respond, in turn 2, that it is Monday, and the teacher evaluates the response in turn 3 by evaluating the answer as correct and repeating the students' earlier response. This brief exchange shows how a teacher uses language to manage and control classroom communication (Johnson, 1995). Outside classrooms, it is unusual to find participants in everyday conversations evaluating responses to solicits; rather, participants usually acknowledge such solicits. For example, using the same example as above in everyday conversations (say in a coffee shop) the response to the question "what day is it?" would be acknowledged (rather than evaluated) with a thank you as follows in episode 2:

Episode 2: Communication Structure Outside Classrooms

Person 1: What day is today?	[**I**nitiation]
Person 2: Monday.	[**R**esponse]
Person 1: Thank you.	[**A**cknowledgement]

The unmarked structure underlying classroom communication distinguishes it from other forms of communication in that teachers usually respond to what their students say, not by replying or acknowledging it, but by evaluating it (Belleck et., al., 1966). In fact, research by Belleck et., al. (1966) revealed that nearly one-third of all teachers' moves consisted of evaluating their students' responses and that many teachers were unaware of this. If, however, a teacher's classroom communication continually follows the unmarked IRE structure, it can be problematic. If teachers constantly evaluate their students' responses, the students may interpret these evaluations as devaluating their opinions, and, in the worst case, an attempt to silence them. This undermines students' ability to contribute successfully to the lesson (Barnes, 1976). Classroom communication must allow for variability in the underlying communication structures if lessons are to be effective (Mehan, 1979; Farrell, 2004b). As Barnes (1976, p. 18) suggests, "learning is not just a matter of sitting there waiting to be taught." Students also have beliefs about learning a language and their own purposes for learning.

Variability in Classroom Communication

Teachers have other methods of using language to control (or give up control) patterns of communication in their classrooms. These, however, are an exception rather than the rule and are, therefore, called marked (unusual) patterns of classroom communication. They can take such forms as student initiations and teacher responses followed by student evaluations, or, students can perform all three moves (Mehan, 1979). Episode 3 below outlines an example of variability in communication structures in a private pre-school English language class in Singapore. In this class, the students were encouraged to initiate and 'take the floor' whenever they thought it correct to do so. The teacher's stated objectives in this phonics lesson were that the students would be able to recognize and read 'nd', 'ng' and 'nk' in words. Episode 3 occurred midway in the lesson (turns 204 to 215) and outlines an exchange in which certain students became curious about the meaning of some words they were pronouncing, especially the word "plank".

Episode 3: Student Initiation

Turns 204-215

204 **Jazin**: What is a plank?

205 **Teacher**: What is a plank? This is a plank. A wooden plank. Can you see...you go to the construction site or you go outside and you see people building something…they would need wood right? And the wood is cut in this shape. When the wood is cut in this shape, this is called a ..?

206 **Sean**: I see this before. My father used this before.

207 **Teacher**: He used this for what?

208 **Sean**: To knock on the bridge...to fix something.

209 **Teacher**: To fix something...some people cut the wooden thing into smaller pieces and maybe they make something.

210 **Jazin**: Or, they cut the tree and then cut, then they make into wooden boxes or this, like that?

211 **Teacher**: That's right. Yes…from the tree they get the plank...they put the...they cut down the plant, whole tree..they put the plant in the machine..

212 **Sean**: Actually on top, they have the…the..?

213 **Teacher**: The tree bark. That's right, they take out the tree bark..They put through the machine and they get planks.

214 **Sean**: They put on my…

215 **Teacher**: You..you see the bookshelf behind? Okay, you see the bookshelf where you put your water bottles?

216 **Sean**: um..hmm.

217 **Teacher**: What is that made of?

218 **Sean**: Planks.

219 **Teacher**: Planks. Wooden planks. That's right.

220 **Jazin**: And they paint it.

221 **Teacher**: Yes..What colour do they paint it?

222 **Sean**: White!

223 **Teacher**: White! That's right..!

224 **Yushan**: That thing...you put books also?

The exchange in Episode 3 clearly shows that during this phase of the lesson, the students took control of communications by choosing the topic for discussion--the meaning of the word "plank." Jazin initiated the question about the meaning of the word after they all had spent time pronouncing it (turn 204). Then the teacher took up Jazin's initiation by responding to his query (turn 205), which was followed by another student, Sean (turn 206). Even though the students initiated and the teacher responded, it does not mean that the teacher had lost control of her classroom communications. Rather, the teacher had full control of the communication in her class, and she was comfortable with her students' taking control of classroom topics. After class, she mentioned that she always hopes that her students are "engaged in constructive discussions that are meaning-focused." Consequently, the students in this class can feel that their teacher values their input, and, as a result, they are not afraid to initiate talking during the class, thus taking responsibility for their own learning.

However, not all student initiations are considered by teachers to be useful or appropriate, and, so, teachers must use other methods of controlling classroom communication to indicate when such student initiations are not welcome (Farrell, 2004b). One strategy teachers use when they are confronted with students who ask questions that are unrelated to the lesson objective (academic task) is to try to ignore the question. This is a teacher's way of keeping control of the topic as outlined in Episode 4 below (from the same phonics class outlined above).

Episode 4: Teacher Ignores Student Initiation

230 **Jazin**: Why Xu never come today?

231 **Teacher**: 'C-and-le'?

232: **Students:** Candle

The brief exchange in Episode 4 shows that Jazin initiated a question (turn 230) inquiring about the nonattendance of a classmate; the teacher ignored this by following with another initiation (turn 231), and then by the whole class response (turn 232). Thus, the teacher ignored Jazin's question as it disrupted the class, and the students recognized this pattern because they responded to the teacher's initiation.

The following episodes (from Farrell, 2004b) further illustrate how teachers vary the underlying communication structure in their classrooms. Episodes 5 and 6 show exchanges from an English language class on grammar that focused on singular and plural nouns. According to the teacher, at the end of the lesson, the "students should be able to recognize the singular and plural forms of nouns, convert the nouns from singular to plural forms or vice versa and choose the correct form for sentences" (Farrell, 2004b:16).

Episode 5: Evaluation of Correct Answer

Turns 57 – 59

57: **Teacher:** When we say 'b' 'b', 'e', e'…what is the last letter there?

58: **Ss**: 'e'!

59: **Teacher:** 'e'! .

Turns 71 – 73

71: **Teacher:** When it's an 'x', what must you put?

72: **Ss**: 'x' 'es'!

73: **Teacher:** 'es'! To show…the plural…

Key: Ss represents students' choral response.

The exchanges in Episode 5 show how the teacher utilized an alternative means of evaluating students' correct responses: she repeated and/or rephrased their correct responses as in turns 57 to 59 and turns 71 to 73. Furthermore, the students interpreted this as an indication that their answers were correct.

Episode 6: Evaluation Incorrect Answer

Turns 24 – 35

24: **Teacher:** Besides people. Besides people, anymore?

25: **Students**: Things!

26: **Teacher:** What else?

27: **Students**: Boxes!

28: **Teacher:** Huh?

29: **Students**: Boxes. Three boxes.

30: **Teacher:** But I'm asking you...here the noun, what does,

what do they refer to, beside people?

31: **Students**: Car.

32: **Teacher:** Cars? What are cars?

33: **Students**: What are cars? (Laughter) Car ah, hah?

34: **Students**: Object

35: **Teacher:** Yes, Object. Thing.

The exchange in Episode 6 shows how the teacher used a different method of indicating to students that they responded incorrectly: she followed up incorrect responses with another initiation, thus indicating that their responses were incorrect. In other words, the teacher did not overtly evaluate incorrect responses. Instead, she followed with a second initiation, which actually acts as an evaluation to 'tell' the students that their responses were incorrect. At the same time, this initiation acting as an evaluation (or I acting as E) indicates to the other students that they are free to respond. This seems to be a common pattern for teachers who do not want to evaluate a student's incorrect answer negatively. For example, in turn 24, the teacher asked a question (I) and the students answered incorrectly in turn 25 (I). Then the teacher ignored this incorrect answer by making another initiation in turn 26. This second initiation, in fact, acts as an evaluation (Farrell, 2004b), or [I *as* E] and the students interpreted this to mean that they must continue to seek the correct answer because they were incorrect in turns 30 (I as E) and 32 (I as E) in which the teacher again made initiations that acted as evaluations, indicating to the

students that they had not given the correct answer. The students finally gave the correct answer in turn 35 and were "rewarded" with a positive overt evaluation.

The exchange outlined in Episode 6 also illustrates how teachers can unwittingly badger their students into submitting the 'correct' answer (the answer the teacher wants) without the students fully understanding the concept being discussed. In this particular case, it seems that the teacher was concerned only that the students learned the grammar terminology and prove this by using it in the required answer ('Object', 'Thing'); however, the students wanted to provide the specific meaningful linguistic examples ('boxes' and 'cars'). As the students' responses were not exactly what the teacher had in mind, she did not accept their answers. Instead, she either asked another question (Turn 26) or recast their responses into questions (Turn 32) so as to try to elicit the answer she wanted. Finally, the students provided the 'correct' answer and then she moved on in the lesson. However, teachers should heed Cazden's (1988) caution that just because the students respond with the correct answer, this does not mean that they have understood the concept being discussed.

Implications for Language Teachers

The key to language teachers' understanding the importance of classroom communication and how this either sets up or blocks opportunities for their students' learning only takes on real meaning when teachers themselves investigate and reflect on the communication patterns in their classrooms (Farrell, 2004b). That is, teachers must gather concrete data about the communications that exist in their classrooms and then use the information garnered from this data to make informed decisions about their teaching (Farrell, 2004a). The most important type of data a teacher should obtain is in the form of classroom transcripts. The teacher collects this type of data by placing a tape recorder and/or video recorder in his/her classroom. Once the data have been collected, the teacher then needs to transcribe the recording, which can be the most painful part of the whole process as it can take up to eight hours to transcribe one class hour class. However, it may not be necessary to transcribe the entire recording as teachers can decide what aspect of the classroom communications they are interested in knowing more about. For example, teachers may be interested only in reflecting on the impact of their verbal instructions in their classes, so all they need to do is listen and transcribe those parts in the tape that show the teacher giving instructions and then the immediate turns after this (for about five minutes) to see the impact these instructions have had on their students' learning. Other topics could include the type and frequency of teacher (and student) questions, how tasks are set up in their classes, or the type of language used in group discussions (for more details on the topics teachers can reflect on in their classrooms, see Farrell 2004b).

After transcribing the classroom communication, the teacher can then analyze and interpret the data. When analyzing classroom communication data, language teachers can choose from three methodological frameworks, each taking a different stance to analyzing and interpreting classroom discourse. These are *Ethnography of Communication* (EC), *Conversation Analysis* (CA), and *Systematic Functional Linguistics* (SFL). It is beyond the scope of this paper to give details of all three

(interested readers should consult Christie, 2002 for a full discussion of these). The essential concepts central to EC are culture, community and communication (Duff, 2002); while CA interprets how turns and sequences of talk are developed in a moment-moment manner and the resources participants make use of to construct and allocate turns (Mori, 2002). Different from EC and CA is SFL, which takes a functional grammar approach to analyzing classroom communication (Christie, 2002). After making interpretations about their classes, teachers can reflect further and decide how they want to make changes (if any) in their teaching. In this way, teachers can take more responsibility for the decisions they make in their classes -- decisions that are informed from concrete classroom data -- not just based on feelings or impulse.

Conclusion

Research evidence has suggested that in many classrooms communication is generally structured to follow a three-part sequence of acts as follows: a teacher *initiation* act (I), followed by a student (or students) *response* act (R), in turn, followed up with an *evaluation* act (E). That said, this paper has also indicated that variations to this unmarked underlying classroom communication structure may exist. These variations are dependent on the purposes of the talk, the number of participants and the medium of interaction (Mehan, 1979). Language teachers must be aware of the variations of classroom communications so that they can reflect on achieving optimum levels of learning in their classrooms. One way of investigating the communication patterns that exist in their classrooms is by recording, transcribing, coding, analyzing, and interpreting these communications. In this way, teachers can become more aware of how their students use language, how they interact with the teacher and each other and how communication patterns in their classroom either sets up, or blocks, opportunities for their students to learn.

References

Barnes, D. (1976). *From communication to curriculum*. Middlesex: Penguin.

Belleck, A., Kiebard, H., Hyman, R. & Smith, F. (1966). *The language of the classroom.* New York: Teachers College Press.

Cazden, C. (1988). *Classroom discourse*: *the language of teaching and learning*. Portsmouth, NH: Heineman.

Christie, F. (2002). *Classroom discourse analysis: a functional perspective.* London: Continuum. Reading, MA: Addison-Wesley.

Duff, P. A. (2002). The discursive co-construction of knowledge, identity and difference: An ethnography of communication in the high school mainstream. *Applied Linguistics,* 23,(3): 289-322.

Farrell, T. S. C. (2004a). *Reflective practice in action*. Thousand Oaks, CA: Corwin Press.

Farrell, T. S. C. (2004b). *Reflecting on classroom communication in Asia*. Singapore: Longman.

Johnson, K. (1995). *Understanding communication in second language classrooms*. New York: Cambridge University Press.

Mehan, H. (1979). *Learning lessons: social organization in the classroom.* Cambridge: Cambridge University Press.

Mori, J. (2002). Task design, plan and development of talk-in-interaction: An analysis of small group activity in a Japanese language classroom. *Applied Linguistics,* 23,(3): 322-347.

Richards, J. C and Lockhart, C. (1994). *Reflective teaching in second language classrooms*. New York: Cambridge University Press.

The Case for Computer-assisted Extensive Reading

Tom Cobb

Background: An ESL dialogue

In scientific dialogues within applied linguistics, turn-taking can involve delays of a decade or more. An example is a recent contribution entitled *Free voluntary reading: Still a very good idea*, in which Stephen Krashen (2003) criticized the findings of a study I was involved in that called into question the extent of vocabulary acquisition that results from pleasurable, meaning-oriented, private extensive reading (Horst, Cobb, & Meara, 1998, *Beyond a Clockwork Orange*). Our study found that even with all the usual variables of a pre/post-empirical, extensive reading study held, and rather more tightly than usual (e.g., more than in some of Krashen's own studies), the number of new words learned from reading a complete, motivating, level-appropriate book of 20,000 words was not sufficient to be the main or only source of vocabulary growth for a learner to expect to function any time soon in English in an academic or professional setting. (It should be noted at the outset that, while vocabulary growth is just one of the potential benefits of extensive reading (along with fluency, grammar and other areas of growth), is often used as simply the most measurable of the various outcomes.)

Krashen (2003) argued that studies like ours typically underestimate the exent of lexical growth that takes place because words are encountered and re-encountered in the course of extensive reading. Many words and phrases, which do not appear in test results, are learned, but this is because of the crude nature of the testing instruments, which typically have no way of accounting for partial or incremental learning. In fact, he argues that as one reads, words are becoming known under the surface and may emerge as new vocabulary only some time in the future.

Over this period, Krashen's views remain largely unmodified—he remains convinced of the value of extensive reading, yet he has never really been able to prove his case, which ultimately rests on a sort of faith. Our views, on the other hand, have undergone some modification, and, in some ways, have come more closely into line with Krashen's own and may even provide these with a firmer foundation than he has provided himself. A number of studies by Horst (2000) and Horst and Meara (1999) have investigated incremental vocabulary growth from reading using a matrix model. This model borrows the notion of vocabulary knowledge as a scale (including points such as, I do not know this word, I have seen this word, I think I know this word, I know and can use this word in a sentence), specifically the vocabulary knowledge scale (VKS) as developed by Wesche and Paribakht (1996). The matrix model is a longitudinal version of this scale that makes it possible to track a word through the knowledge levels over time as it is encountered and re-encountered.

Part I: Computing the vocabulary learning from extensive reading

The matrix uses numbers from 0 to 3 to indicate the points on the scale, as follows:

0 = I definitely don't know what this word means
1 = I am not really sure what this word means
2 = I think I know what this word means
3 = I definitely know what this word means

These numbers are then placed on a simple two-dimensional graph, with the same numbers appearing both top to bottom and left to right, as can be seen in Figure 1.

	0	***1***	***2***	***3***
0				
1			x	
2				
3				

Figure 1: From scale to matrix.

Every cell in the matrix is an intersection between two numbers; i.e., 'x' is at the intersection of 1 and 2. This means that a word was rated as a 1 ('I'm not sure') after a previous reading encounter, but then was rated as a 2 ('I think I know') after a subsequent encounter. In other words, cell intersections represent partial word learning as a result of a further encounter with a word. The movement between 1 and 2, or 2 and 3, represents an increase in knowledge of the word, but not enough of an increase to register on most vocabulary tests.

Employing a methodology of repeated readings of a literary novella and a computer-based testing apparatus that allows us to test large number of words in a relatively short time, we have been able to trace the ups and downs of word knowledge that normally passes below the radar of conventional tests. In one study, Horst (2000) tracked 300 words through several readings of a German novella and, after each reading, identified how many words were at each knowledge level as compared to the previous reading. Each additional reading produced a new matrix with the 300 words distributed slightly differently over its 16 cells each time. Authentic data for a pair of readings from this study, as reported in Horst (2000), is shown in Figure 2.

	0	***1***	***2***	***3***
0	**75**	27	9	3
1	4	**20**	20	6
2	2	4	**13**	35
3	0	0	7	**75**

Figure 2: Movement between readings

The numbers in the cells refer to the number of words inhabiting each intersection point, with the row label indicating the previous knowledge level for those words and the

column label indicating current knowledge. The bold numbers on the diagonal represent the number of words that had not moved between readings (75 were rated 0, unknown, in the previous reading and were still rated unknown following a second reading, etc.). Notice that words to the left of the diagonal.have lost ground since the previous reading and are rated as less known than previously, while words to the right have gained ground.

Through simple addition, one can see that there are more words above the diagonal (27 + 9 + 3 + 20 + 6 + 35 = 100) than below it (4 + 2 + 4 + 7 = 17), and, hence, that more are gaining than losing ground. Over the several readings and matrix calculations of Horst's (2000) study, it became clear that it was indeed the case that words were generally making progress over the course of several readings much of which would, nonetheless, not have registered on a standard vocabulary test with an all-or-nothing assumption about word knowledge (such as Nation's classic, 1990, *Vocabulary Levels Test*, or Laufer & Nation's, 1999, update). Only 44 of Figure 2's words (3 + 6 + 35 = 44) had moved into "I know this word," but another 56 had made lesser gains. To summarize, then, while the implications of this methodology are still being worked out and will appear in forthcoming reports, it already seems clear enough that Krashen is right: there is more word learning from extensive reading than meets the eye.

This modification of our views, however, is in some ways rather minor, inasmuch as we still do not consider that extensive reading as traditionally practised could ever be the only or even main source of vocabulary growth for a second language learner. That is because however incremental the learning of encountered words may be, it still requires that words *be* encountered in sufficient number, and simple corpus research makes it clear that this will not happen.

How many words are enough to begin a serious undertaking in a second language, such as academic study or professional activity? Vocabulary researchers like Laufer (1992) and Hirsch and Nation (1992) tag the number at a minimum of 3000 word families, provided these are carefully selected for frequency and text coverage. It has also been shown that between six and ten encounters are needed for learning to occur (Zahar, Cobb & Spada, 2001), and in our own matrix work it seems that at least six encounters are needed for a word to travel reliably from rating 0 to 4 and stabilize there. Will 3000 word families be met six times through extensive reading?

A computer program called *Range* (developed by Heatley & Nation, 1994; adapted for Internet by Cobb; available at the *Compleat Lexical Tutor* website, http://www.lextutor.ca; shown in Figures 3 & 4) takes a user's word or expression as input and determines how often it occurs in a broad corpus of English writing. This corpus (the Brown corpus, Kucera & Francis, 1979) is divided into 15 sublists, from science to romantic fiction to law. For the purposes of the present argument, the Brown corpus can represent the most an ESL reader could possibly read in a year or two of extensive reading—most learners would obviously read both much less and much less broadly. The surprise finding is this: after the most frequent 1000 words of English, words thin out quite rapidly. Here are some figures from the 2000 and Academic Word

List components, which Nation and others include in the necessary lexicon of 3000 word families.

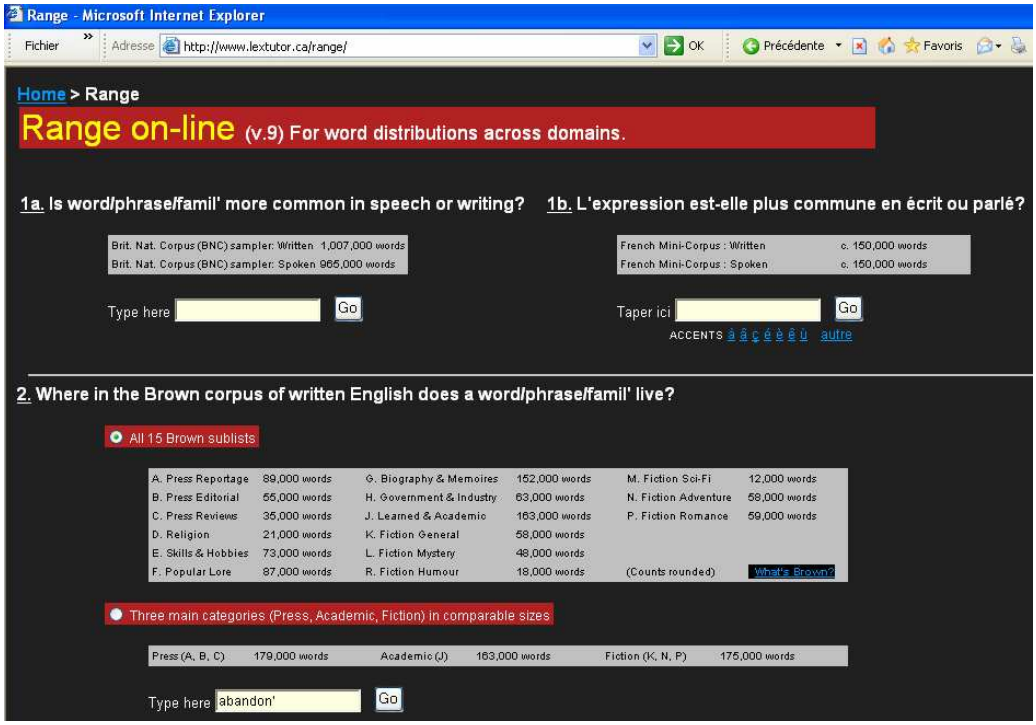

Figure 3: Range for word distributions – input

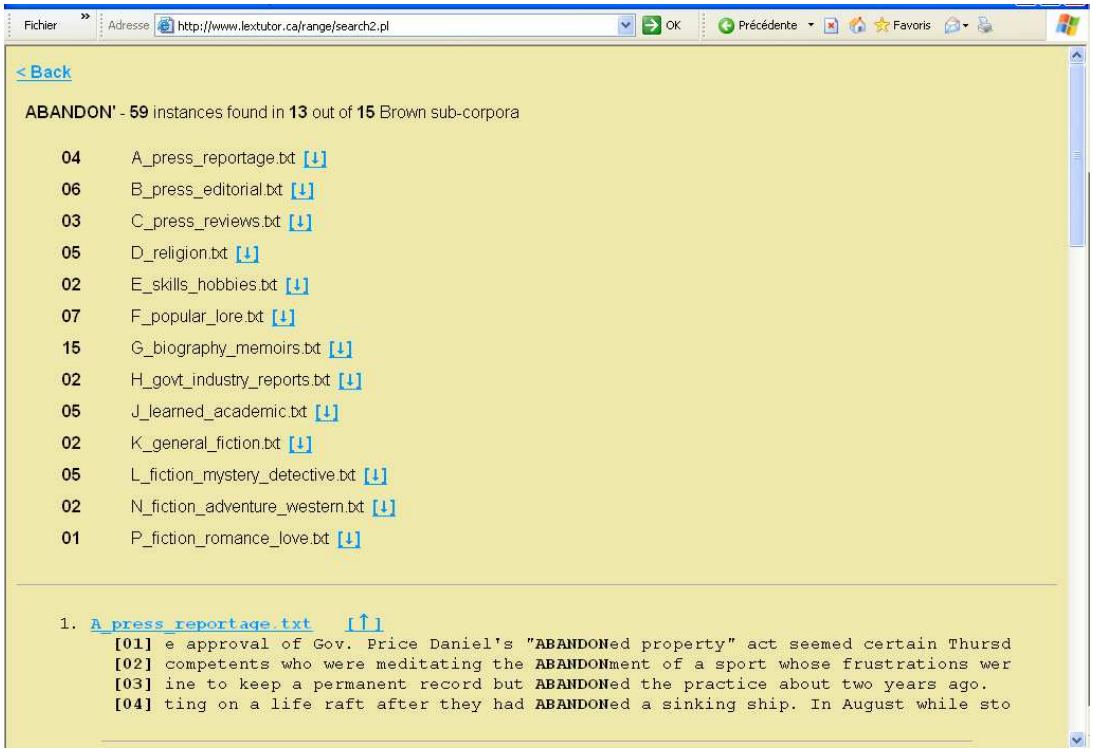

Figure 4: Range for word distributions – output

The distribution of the word family *abandon* throughout the Brown corpus is requested in Figure 3 and shown in Figure 4. The point to notice is that, while the item appears in most of the sub-corpora, 13 out of 15, it appears more than six times in only three of them. This pattern is general for even medium-frequency words. Table 1 shows the distributions in the 15 Brown sub-corpora of six word families from the high frequency 1000 list (not including function words), while Tables 2 and 3 show distributions for six word families from higher up in the 2000 list and another six families from the academic word list (AWL). Readers can visit Lextutor for themselves and enter their own words into the analysis program.

1000-level word family	occurrences in 1 million words	Present in how many from 15 sub-corpora	in how many with 6+ occurrences
car	285	13	10
house	760	15	14
country	510	15	15
able	216	15	15
add	715	15	15
admit	473	15	15
Mean	493.2	14.7	14
S.D.	200.5	.7	1.8

Table 1: Word distributions for high frequency words

2000-level word family	Occurrences in 1 million words	Present in how many of 15 sub-corpora?	In how many with 6+ occurrences?
accuse	46	9	3
accustom	15	10	0
ache	4	3	0
admire	10	8	0
afford	58	12	4
alike	20	10	0
Mean	25.5	8.7	1.2
S.D	19.7	2.8	1.7

Table 2: Word distributions for medium-frequency words

AWL word family	Occurrences in 1 million words	Present in how many of 15 sub-corpora?	In how many with 6+ occurrences?
abandon	59	13	3
academic	95	13	6
accumulate	29	6	3
achieve	223	13	8
acknowledge	32	11	1
acquire	100	13	6
Mean	89.7	11.5	4.5
S.D	65.6	2.6	2.4

Table 3: Word distributions for AWL words

We can now answer the question: will 3000-word families be met six times through extensive reading? If the Brown corpus represents the language at large, as it was designed to do, then it seems clear that an extensive reader following his or her interests through one or two text domains or sub-corpora would meet the most frequent 1000 words in great abundance in any domain, but would meet even slightly less frequently appearing words only intermittently and probably not often enough for learning to occur. A word as common as *ache* would not be met six times in any of the Brown's 15 sub-corpora; a word as common as *accumulate* would be met six times in only three of the Brown's sub-corpora. In other words, meeting any significant portion of the critical 3000 words, six times each, in extensive reading, is rather unlikely.

In summary, while there may be more word learning than meets the eye from random encounters in extensive reading, as Krashen believes, nevertheless most words will simply not be encountered. The distribution of words in English simply does not allow a sufficient number of encounters to take place in this manner.

Part II: Bringing up the numbers with computing

There are a number of teaching strategies that can increase the odds of words being encountered and learned. Some of these involve the "direct teaching" of vocabulary, for example, through the classroom use of word lists or supplementary vocabulary course books, and, of course, through *ad hoc* teacher attention to vocabulary queries in class. But there are also strategies that can be devised to increase the odds for extensive reading, itself -- many involving recent developments in computer technology (assuming that the extensive reading materials are in machine-readable format, which is increasingly the case). The rest of this paper will outline several of these. The means for implementing these strategies are available to teachers or researchers on the website mentioned above. The format of the following discussion will adopt a framework of problem, solution and research on the solution as well as proposals for distributing the solution to others in the form of 'builder' or authoring systems.

Problem 1: The number of encounters with new words is lower than it needs to be because learners do not always recognize a word they have met in text when they meet it again in speech.

It is common for many advanced learners to have an extensive lexicon of medium- and lower-frequency items for which they have only weak or uncertain sound representations. For this reason, if they re-encounter in speech a word previously met in reading, they often do not recognize it as a second encounter. How big a difference could it make if learners knew the pronunciation for every word they met in reading? Corpus evidence can provide an idea.

It is well established that conversational English comprises mainly (about 90%) 1000-level word families, and, conversely, that post-1000 items are mainly to be found in texts of various types (Stanovich & Cunningham, 1992). But this does not mean that less

frequent words are totally absent from spoken English. The same sample of words referred to above were put through another of Range's distribution comparison routines -- this time one comparing similar-sized corpora (roughly 1 million words each) of spoken and written British English as found on the BNC corpus sampler collection. Tables 4 and 5 show the pattern of these distributions for 2000 and AWL-level medium-frequency words.

2000-level word family	Occurrences in 1 million words of writing	Occurrences in 1 million words of speech
accuse	64	7
accustom	42	2
ache	5	5
admire	61	14
afford	43	78
alike	20	4
SUM	235	110
MEAN	39.17	18.33
SD	23.03	29.52

Table 4: Speech vs. writing for medium-frequency words

AWL-level word family	Occurrences in 1 million words of writing	Occurrences in 1 million words of speech
abandon	45	6
academic	81	6
accumulate	30	11
achieve	199	91
acknowledge	34	14
acquire	158	10
SUM	547	138
MEAN	91.17	23.00
SD	71.19	33.45

Table 5: Speech vs. writing for AWL words

These tables show that post-1000 items are, indeed, found a great deal more in text than in talk -- more than twice as much for 2000-level words and nearly four times as much in the case of AWL words. From another point of view, however, one can say that if learners knew that the sound of every word they had met and noticed in reading, then they could increase the number of occurrences of new words, (in the sense of recognizing them as re-occurrences when they re-encountered them in speech), by as much as 46% (110/235 x 100) for 2000-level words and 25% (138/547 x 100) for AWL words. Of course, as in the case of the Brown corpus, above, these occurrences may well be unevenly distributed within the BNC sampler corpora, but these are not broken into sub-

corpora; therefore, we cannot easily find out. For instance, *acquire* in Table 5 could well be piled up largely in the second language 'acquisition' corner of the corpus, so that engineering students might be unlikely to encounter this item.

Some distributional information for medium-frequency words in speech, at least as employed in academic contexts, can be gathered from the University of Michigan's MICASE (Michigan Corpus of academic spoken English) corpus and website (Simpson, Briggs, Ovens, & Swales, 2002). This corpus is almost double the size of those consulted above, at 1,848,364 words broken down across several topics, situation and speaker types as well as domains (although, unfortunately, not across the same domains so that the comparison is rough rather than precise). Table 6 shows a small sample of AWL words as broken down by topic areas. Again, the reader is welcome to expand the sample by visiting the site and entering other words, at http://www.hti.umich.edu/m/micase/.

AWL-level word family	Occurrences in 1.8 million (m) words of speech in academic contexts (MICASE)					
	Biology-Health Sciences	Arts, Humanities	Social Sciences, Education	Physical Sciences, Engineering	Occurrences in 1.8 m	Occur. in BNC spoken, 1m
abandon	4	12	3	36	55	6
accumulate	29	2	21	22	74	11
achieve	10	13	59	11	93	91
acknowledge	7	20	7	4	38	14
acquire	4	27	16	26	73	10
SUM	54	74	106	54	333	132
MEAN	10.80	14.80	21.20	10.80	66.60	26.40
SD	10.47	9.36	22.30	10.47	20.89	36.23

Table 6: AWL items in academic speech vs. general speech

Two points emerge from this brief look at the MICASE data. The first is that, again, encounters are likely to be piled up unpredictably rather than distributed evenly (as *acknowledge* is piled up mainly in Arts and Humanities in Table 6). The second, nonetheless, is that if a learner was learning English in order to function within an academic environment, then spoken language within this environment yields a somewhat higher proportion of post-1000 items than does the spoken language generally. *Accumulate* appears 74 times in 1.8 million words of spoken academic, as against 11 times in 1 million words of general English (or, one can extrapolate, 11 x 1.8 = 20 times in a general speech corpus of equal size). Lesser, but still substantial, advantages for the academic corpus are shown for *abandon, accumulate, acknowledge* and *acquire* (but not for *achieve*). To conclude, it seems safe to say that the 25% increase in occurrences shown above for knowing how words are pronounced in general spoken English could be somewhat greater within such target domains as academic speech.

How can we ensure that learners have full access to the pronunciation of any new word they happened to come across in their reading? Lextutor's builder routines offer two ways of doing this.

Solution 1: If learners are reading a text on a computer screen and have access to the Internet, then Lextutor gives them the means to access the pronunciation quickly and simply for any word in the text. At one of the website's Hypertext routines (available at http://www.lextutor.ca/hypertext/), learners can enter any text they happen to be reading into a text input and click to transform it into a text with literally *every* word linked to a text-to-speech engine giving a tolerable (or better) pronunciation of the word. This requires a once-only download of a speech plug-in (free from Macromedia), which, once completed, allows instant pronunciations that do not distract readers from their reading unduly.

Solution 2: Of course, there is no guarantee that one audition of a word leads to a stable memory for its contours or a stable link between its phonetic and semantic features. If learners or their teachers wish to ensure a particular set of words is heard again soon, they can make use of Lextutor's Dictator routine (at http://www.lextutor.ca/dictator), which transforms any word list into a text-to-speech based spelling activity, in either practice or test formats. In Figure 5, a learner has created a training exercise to practise spelling the words he or she hears. The learner clicks a word to hear it, tries to spell it and is given help with any errors. The help is provided by the resident Guidespell tutor, first discussed in Cobb (1997a); in this case, Guidespell tells the student how many letters were correct in the attempt to spell *accompany*. When ready, the learner can enter the same words into a Test version of the program, shown in Figure 6, where there is no help but simply a score presented when all words have been entered.

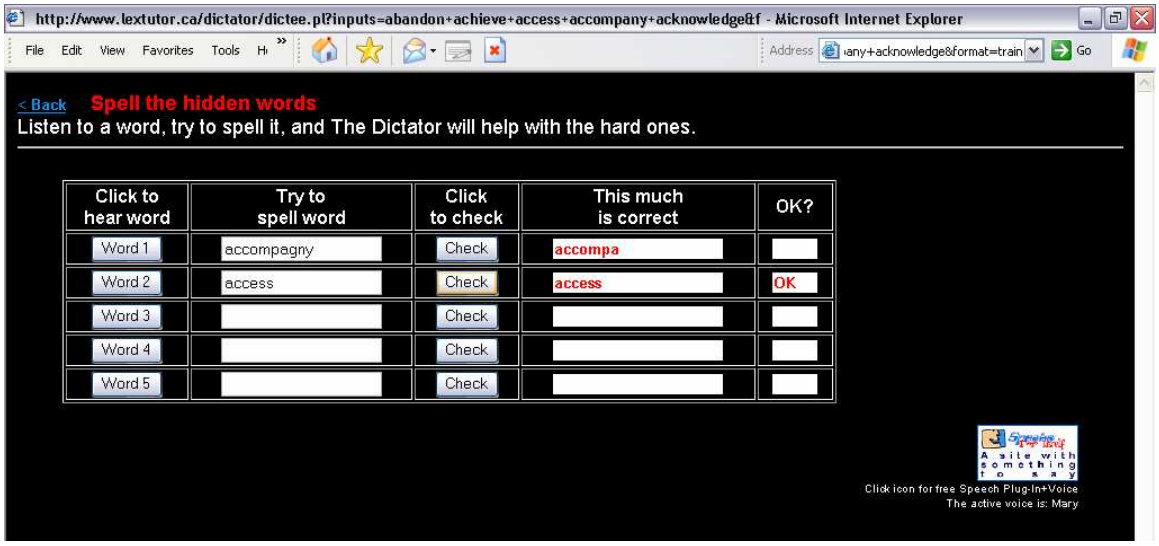

Figure 5: Dictator training activity under way

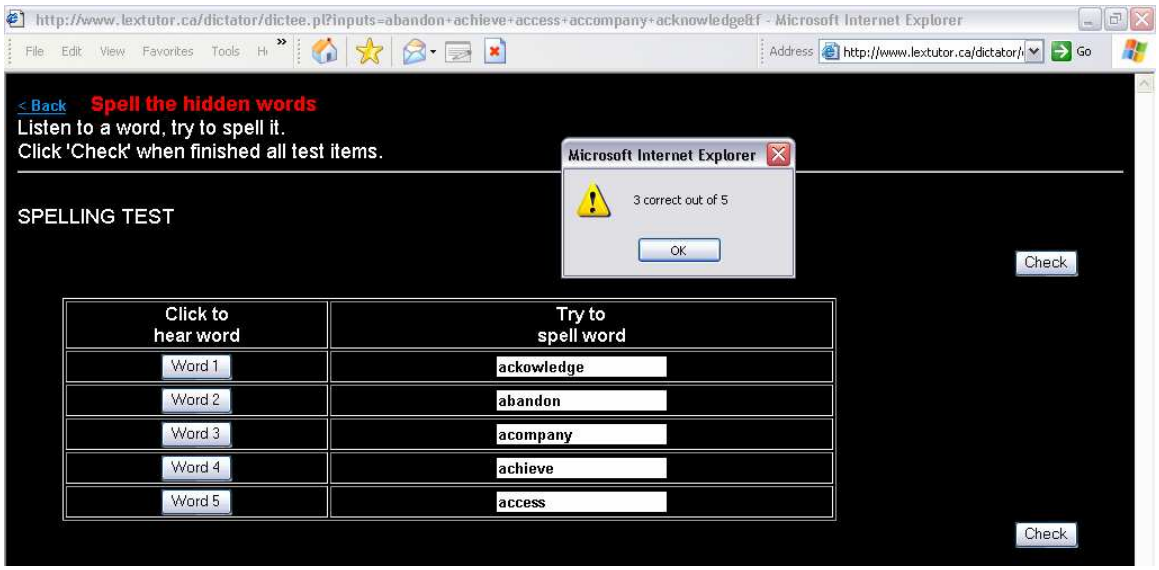

Figure 6: Dictator in test mode.

Hypertext and Dictator, thus, provide two approaches to helping learners form sound-spelling correspondences -- the goal of increasing the likelihood of recognizing words they have met in speech, and, hence, of increasing the likelihood of meeting words enough times to learn them.

Research on Dictator: Realized in December 2004, Dictator has not yet been subjected to substantial empirical testing.
Current work on Dictator (April 2005): Unobtrusive auto-links to Dictator from other reading activities.

Problem 2: Words are often simply forgotten between encounters -- even within the same text.

Research indicates that, on the one hand, new words tend to get ignored if they require a great deal of effort to process them, but, on the other hand, that they also tend to get forgotten if they require very little effort (Mondria & Wit-deBoer, 1991). In other words, it seems the conditions for retention of words from reading are rather particular and may present themselves only occasionally, which may be part of the reason that a minimum of six, and as many as ten, encounters are needed for words to begin to stick. Learners need some way to keep track of whether they have seen a word before and, if possible, to revisit the previous occurrence without a major exit from their current reading. This is something a computer can provide.

Solution 1: One solution is to offer readers a quick way of clicking on a word and recording it for later reflection without losing the thread of the story, as the user shown working in Figure 7 has clicked (with the Alt-key held down) on the words *toil* and *groping* from the first chapter of Jack London's classic tale *Call of the Wild,* located at http://www.lextutor.ca/CallWild.

Figure 7: Tagging *toil* and *groping* for post-reading attention

Solution 2: Another solution is to link a text to a Story Concordance. A reader clicks on any word in the reading text and receives a full accounting of all the occurrences of the same word as already seen or yet to be seen in other parts of the story. The reader in Fig. 7 has just found out that the rather odd word *toil* might be worth paying attention to as it occurs in five of the chapters to come.

Research on Story Concordance: This is another new program that, as yet, has not been extensively tested (although an earlier version of this concept was tested in Cobb, Greaves & Horst, 2001). Nor is there a Builder version of the program -- owing to the complexity of taking user-text input, consisting of extensive texts and divided into several chapters or sections to provide the sort of output seen in Fig 7.

Current development on Story Concordance: The 'Wordbox' is now linked to Dictator, discussed earlier, so that once the learner has collected a few words, these can be sent directly to a spelling activity. This linkage is shown in Figure 8. This feature will become part of a Builder Hypertext routine in 2005, so that teachers and learners can link their own texts to this resource.

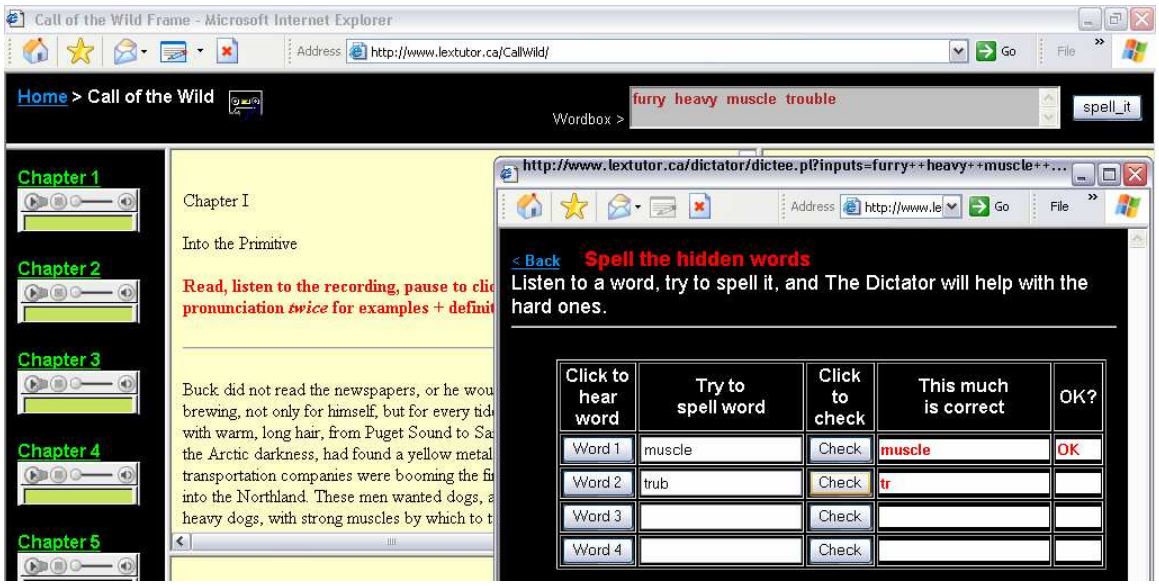

Figure 8: Linking resources

Problem 3: The semantic features comprising word meanings are distributed over several occurrences of words, so that integrated meanings are difficult to construct.

As already noted, post-1000 words are distributed thinly in natural language, but the semantic features comprising the meanings of these words are distributed even more thinly. The whole set of semantic features underpinning the concept represented by a word of any complexity is, inevitably, not present in every occurrence. What this means for a learner, building a vocabulary from reading, is that even if words are not forgotten between encounters, there may not be enough information in a single encounter, or even in a number of encounters, to provide more than a partial sense of its meaning.

For instance, the semantic deep structure of a common word like *work* embraces features ranging from doing a job for pay ('work at a store'), the job itself ('it's my work'), expending effort not necessarily for pay ('work on my car'), correct functioning of a device ('it works now'), to instances of high art ('a work of Shakespeare'), but only one or two of these features are present in a sentence like "What are you working on?" The word learner is thus not only required to remember words from occurrence to occurrence, but, at the same time, to be revising, updating and especially integrating hypotheses as to their meaning(s).

Solution 3: One way of showing learners several pieces of a word's meaning all at once so they need not attempt to gather them all up for themselves with attendant forgetting and backtracking, is to present the word in a concordance (a concordance for 'work' from the Brown corpus is shown in Figure 9). Even a relatively small concordance reveals to an observant (or possibly to a trained) learner such information as the main parts of speech for the word (*the work, we'll work, to work*), several of its senses (*take my car to work, work on my kicking, an idea that would not work*), and its main collocations (*work for, work at, work out,* and especially *work on*).

1 after the board of canvassers completes its work. A difference of opinion arose between Mr
2 Authority bonds for rural road construction work. #A REVOLVING FUND# A01 1310 4 The depa
3 ghes Steel Erection Co. contracted to do the work at an impossibly low cost with a bid that wa
4 lta Sigma Pi at Lamar Tech, and did graduate work at Rhodes University in Grahamstown, South A
5 bomb tore his car apart as he left home for work. Battalion Chief Stanton M. Gladden, 42,
6 blic relations director, resigned Tuesday to work for Lt. Gov. Garland Byrd's campaign. A01 1
7 07 1230 7 #MISSIONARY EXPLAINS# "I don't work for the Government", the American said. "I'm
8 0 12 scrimmaged for 45 minutes. "We'll work hard Tuesday, Wednesday and Thursday", Meek
9 home so that he could take his other car to work. "I'd just turned on the ignition when th
10 school teaching certificate. A normal year's work in college is 30 semester hours. A02 1430
11 A. Berger firm, a Philadelphia builder, for work in the project. The second agreement perm
12 rk out about an hour on Saturday, then we'll work Monday and Tuesday of next week, then taper
13 f cars "might not be realistic and would not work". Mrs. Molvar asked again that the board
14 ARTIST# Mrs. Monte Tyson, chairman, says the work of 100 artists well known in the Delaware Va
15 overhauling of 102 joints. The city paid for work on 75, of which no more than 21 were repaire
16 ly involve failure to perform rehabilitation work on expansion joints along the El track. The
17 e. "This year, coach Royal told me if I'd work on my place-kicking he thought he could use
18 ales will begin and contracts let for repair work on some of Georgia's most heavily traveled h
19 - His miracles - His substitutionary work on the cross - His bodily resurrection fr
20 g, said the transit company is reviewing the work on the El. "We want to find out who knew
21 rty, appeared on payment vouchers certifying work on the project. Varani has been fired on cha
22 as completed after nearly eighteen months of work on the question of the organization of the U
23 bly will have a good scrimmage Friday. We'll work out about an hour on Saturday, then we'll wo
24 several more drafts". Salinger said the work President Kennedy, advisers, and members of
25 aborers go home Tuesday night for some rest. Work resumed Wednesday, he said. Mr. Schaefer
26 , stressed the need for the first two years' work. "Surveys show that one out of three Amer
27 e traditional visit to both chambers as they work toward adjournment. Vandiver likely will men

Figure 9: Lines from the Brown corpus for *work*

However, there are at least three major problems with using concordances as an aid to building a second lexicon. First, learners do not usually have a concordancer handy when they are reading; they would have to write the word down with a certain amount of context and look for it in a concordancer later. Second, most full-blown corpora, like the Brown corpus, are likely to include a high proportion of other words the learner will not know in addition to the one they are looking up. Third, the single chopped-off lines of the concordance format, while designed to highlight immediate formal patterns such as grammar and collocation, also reduces the amount of semantic context below what learners may need to work out the meaning of a word.

Figure 10: Making concordance information accessible and comprehensible

Lextutor offers teachers and learners responses to each of these problems -- all of which can be seen in further developments of the *Call of the Wild* Story Concordancer. The first development addresses the access problem. As already seen, any word in the story, when clicked, generates a concordance instantly in a window within the same eye-span or frameset. The learner can, thus, compare several examples of the word, along with the original, at the same time, with minimal exits from the story. The second problem, of unknown items within the concordance, is addressed by the fact that the concordance is recursive (any word clicked in the concordance, itself, generates a new concordance, which may shed light on the unknown word), and that it derives not from a general corpus but, rather, from a collection of other works by the same author. A same-author corpus should mean that the range of lexis and types of contexts is somewhat constrained relative to a general corpus, has extensive re-cycling built in and offers a consistency of tone and style to which learners can habituate themselves. In the screen print shown in Figure 10, the user has clicked the link "*progeny* in other Jack London stories" (not shown) and is presented with uses of this word from other works by the same author such as *White Fang* and *Martin Eden*. The third problem, the small contexts and chopped off lines, is resolved by building in a mouse select-and-release feature, in which the learner selects several words, releases and is delivered a series of much-expanded contexts, either from the original text or from throughout the London opus, depending where the request is launched (as shown for the phrase *helpless progeny* in Figure 11). For truly astute

learners, this same feature allows them to explore an author's trademark collocations and grammar preferences.

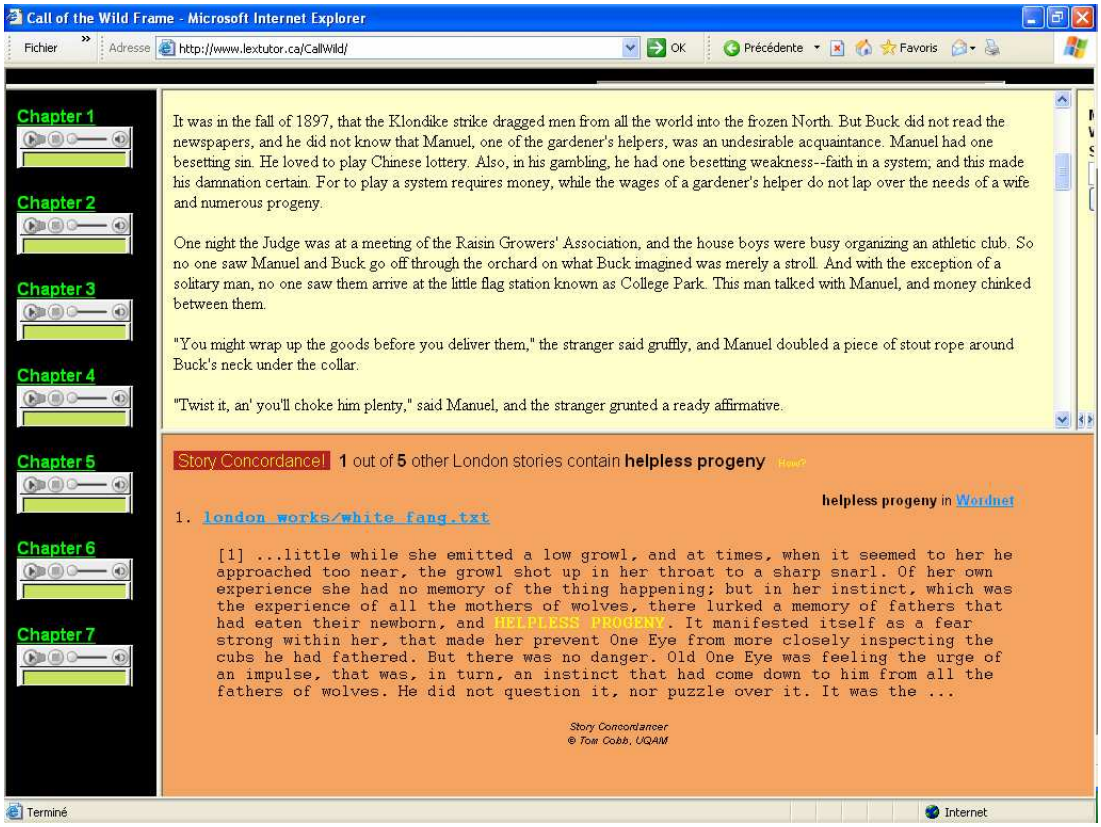

Figure 11: Comprehensibility through same-author corpus

Development - Builder versions of Story Concordance: Lextutor can incorporate a user's text into a suite of reading resources including *most* of those seen on the *Call of the Wild* page, available at http://www.lextutor.ca/hypertext. A screen print, based on a user's text, is shown in Figure 12. Unfortunately, in this routine, it is not currently possible to include a text in separate chapters.

In Figure 12, the corpus accessed by clicking on words is the Brown corpus, which of course has the problems mentioned before. It is not currently possible to allow teachers or learners to load their own corpora into a web-based concordance. However, a number of experiments are under way on Lextutor to allow significant upload of user texts up to 50,000 words (about the size of a Jack London story), including a Text Concordancer (which the reader can inspect at http://www.lextutor.ca/concordancers/text_concord/). Also, more learner-friendly corpora are being developed to replace the Brown in the Hypertext routines, including a corpus of simplified readers that has recently become available.

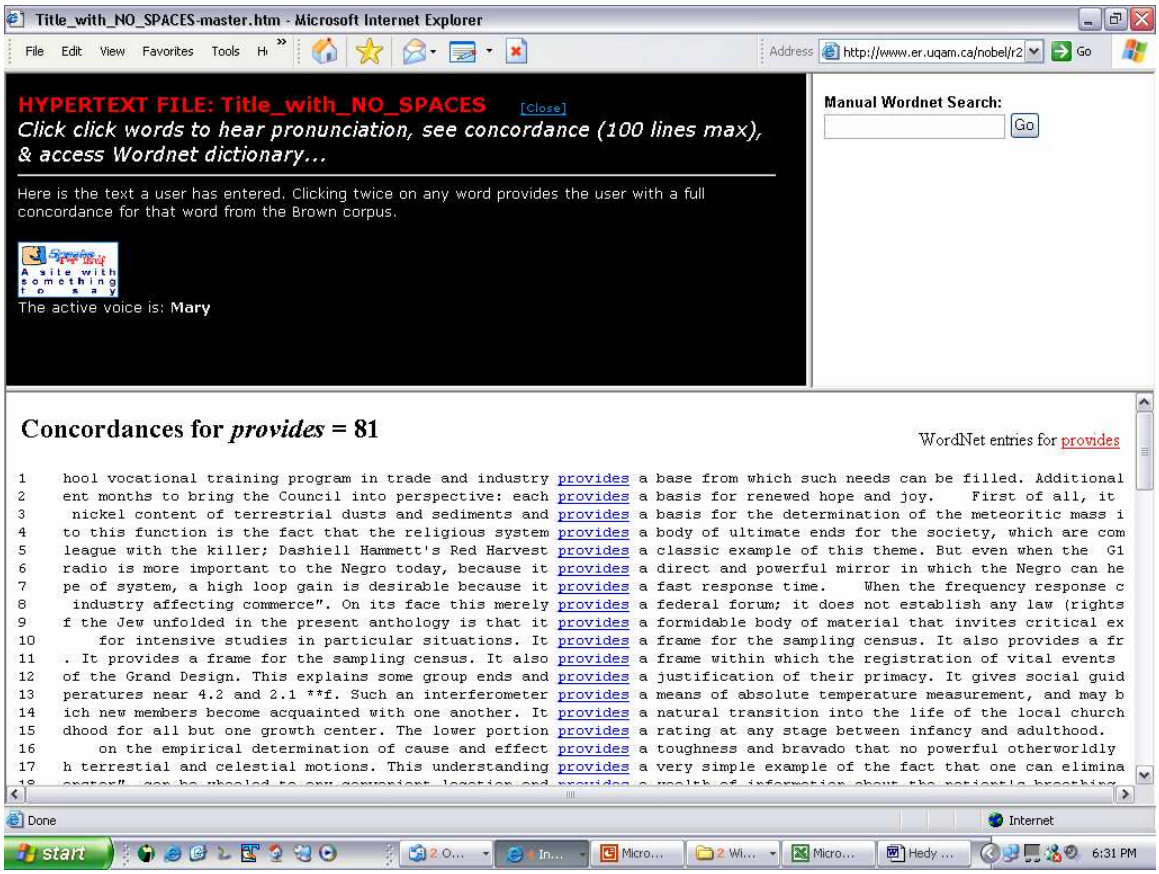

Figure 12: Linking Brown concordances to user-input texts

Research: Does working with concordances, however accessible, however linked to tailor-made corpora, lead learners toward concept integration? In my own research (reported in Cobb 1997a, 1997b; 1999), I proposed that degree of transfer of word knowledge to a novel context should reflect the degree to which a word had achieved a complex semantic representation, inasmuch as a novel context was unlikely to have the exact semantic features present in the word's initial encoding. Subjects in a series of experiments learned words over several weeks using either small bilingual dictionaries or else purpose-built, monolingual concordances. They were then asked to match learned words to short definitions as well as integrate them into a rational cloze passage for a text they had never seen (that embedded the target lexis in contexts made up of familiar words). These tasks are shown in Figure 13. After an extensive training period, students in both control (dictionary) and experimental (concordance) groups had improved equally in their ability to match words to short definitions, but the experimental group had improved significantly more in their ability to apply learned words to novel contexts. These results are shown in the line graphs in Figure 14. (These figures, along with further details, are available online in the author's doctoral study.) This result was replicated a number of times and at a number of levels.

992 YOCABULARY QUIZ 2 (C)

Part A. Recognition (1 each = 15 points)
Write the number of each word beside the correct definition.

1. candle
2. cabbage __conversation
3. candy __green vegetable
4. capital __teachers use it
5. chalk __important city
6. chat

1. centre __two people or two things
2. chapter __part of a book
3. chimney __doctors work there
4. church __Christian building
5. clinic
6. couple

1. cry
2. control __make unhappy sounds
3. charge __make noise with hands
4. cover __ask to pay; cost
5. contact __have power over
6. clap

1. curious
2. cruel __causing pain, hurt
3. crazy __not sunny
4. cloudy __always asking
5. confident questions
6. convenient

Part B. Cloze (1 each = 10 points)
Choose from these words to complete the passage.

continued	customers	catch
century	court	congratulations
cheat	clever	caused
connected	curly	contained

Young Computer Criminals

Computers have changed our lives in many ways and made so many things easier to do. Most of us would probably agree that the computer is the most important invention of the 20th (1)..... By now most of us know how to use a computer and these days children learn how to use them at school. This sounds like a great idea. Computer games are good entertainment and educational programs are clearly very useful. But some children are able to do surprising things with computers. For example, in the United States some young computer users found a way to rob banks! These very (2)..... boys made a computer program that wrote thousands and thousands of different telephone numbers. When they (3)..... their program to the bank's telephone, the computer started dialing the numbers. The computer (4)..... dialing the phone numbers, day after day, until the bank's secret telephone number was finally found. Then the boys were able to open the files which (5)..... bank account information, and they started changing things. As you can imagine, you could easily make yourself very rich just by opening a new file and typing some numbers! Of course, bank (6)..... soon noticed that there were problems with their accounts and they started complaining to the bank manager. The police were brought in but it wasn't easy to (7)..... these criminals because no one expected the thieves to be children. In the end the police weren't sure what to do with the boys. They were too young to be sent to a (8)..... of law, and they were certainly too young to go to prison. The bank just had to find a better way to protect its files. Computer criminals have also (9)..... problems for credit card companies in the same way, and these companies have lost a lot of money. Finally, you may be interested to know that students who are good with computers have found a new way to (10)..... Again, by trying thousands of different numbers, they find their way into the computer network of their school and simply change their marks!

Figure 13: Testing two kinds of lexical knowledge – definitional and contextual

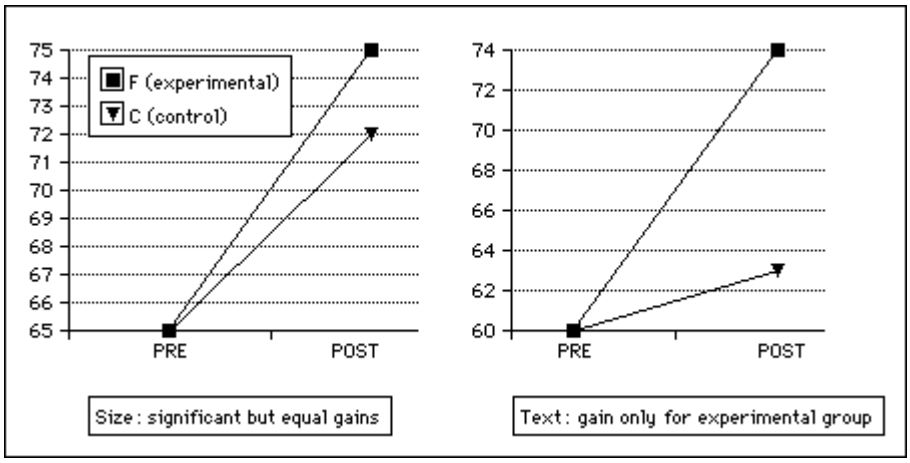

Figure 14: Better transfer to novel texts for concordancers

Problem 5: Beyond the medium-frequency zones, words appear so infrequently in texts that learners have almost no chance of learning a significant portion of them.

As vocabulary acquisition proceeds beyond the 3000-word level, the likelihood of learners meeting many of the remaining 20,000 or so word families known to native speakers becomes very poor indeed. Some learners may not aspire to know all the words that native speakers know, and, for these learners, 3000 may be enough, or, as counselled by Nation and colleagues, it may be time for their efforts to focus on strategy development or reading within an academic or professional domain (Nation, 2001). However, many learners do aspire to full membership in a second community or culture, such as many non-English speaking or non-French-speaking immigrants to Canada, and for these learners, vocabulary growth is a slow and haphazard process.

Figure 15: Sharing new acquisitions.

Solution: Advanced vocabulary acquisition is normally a solitary process, but it need not be. In a class of 20 learners, if each one meets 50 words in a month of extensive reading, then that would amount to 1000 words (possibly with some redundancy) for the group as a whole. Networked computing should, in principle, make it possible for such a group to share lexical acquisitions, while, at the same time, provide for further encounters, retrievals and novel contextualizations, in line with points raised above. Such is the goal of the Group Lex Database at http://www.lextutor.ca/group_lex/demo, a set of web pages allowing learners to enter words from their reading, share words with others, quiz themselves on some or all of the words and quiz themselves with the same words in novel contexts. Figure 15 shows the words as initially entered (in this case, by random visitors to Lextutor). Several areas on the screen shown are hyperlinked to various sorting and

extraction options; for example, clicking on a name will extract all the entries for that name, or, similarly, for a subject area like 'Arts' or other groupings. The quiz option allows a user to select several words for retrieval practice, as shown in Figure 16. This retrieval is, of course, within the original context, but a click on the 'Tougher Quiz' option takes quiz-takers to a new task (Figure 17) that asks them to plug these same words into gapped concordance lines from the Brown corpus – i.e., to transfer their meanings to a novel context (to re-visit a theme from above).

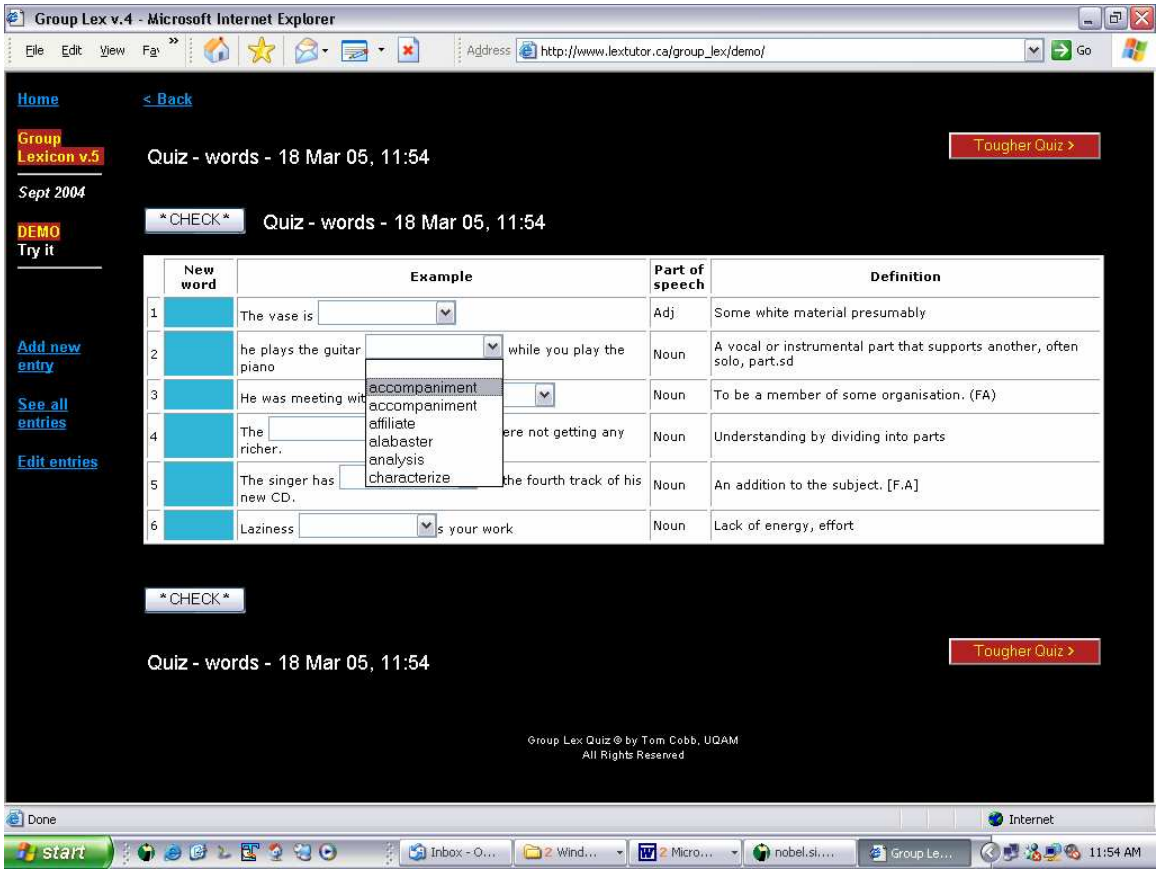

Figure 16: Learner-designed, collaborative instruction

Builder versions: The complexity of these multi-page, database-oriented programs has, until now, delayed the development of Builder (i.e. user-produced) versions of Group Lex. However, several dedicated versions have been set up for roughly 25 teachers in 2004-2005.

Research: Some initial research on learner use of Group Lex is reported in a paper in *Language Learning and Technology* (Horst, Cobb & Nicolae, 2005). Questions, so far investigated, include learning effects, resource-use preference and ability of learners at different levels to generate contexts and definitions that their peers can use and make sense of in the quizzes. Also, the paper by Horst, in the current volume, will discuss further investigations currently under way.

Further development: First, programming is almost completed to connect Group Lex directly to learner texts; for example, the dictionaries are linked to texts in examples above. A learner will select an example sentence containing a target word, which on mouse release, will be sent to an input form for Group Lex. Second, code is being developed to allow teachers to control auto-archiving of a word set when it has reached a certain size or a text is completed, etc.

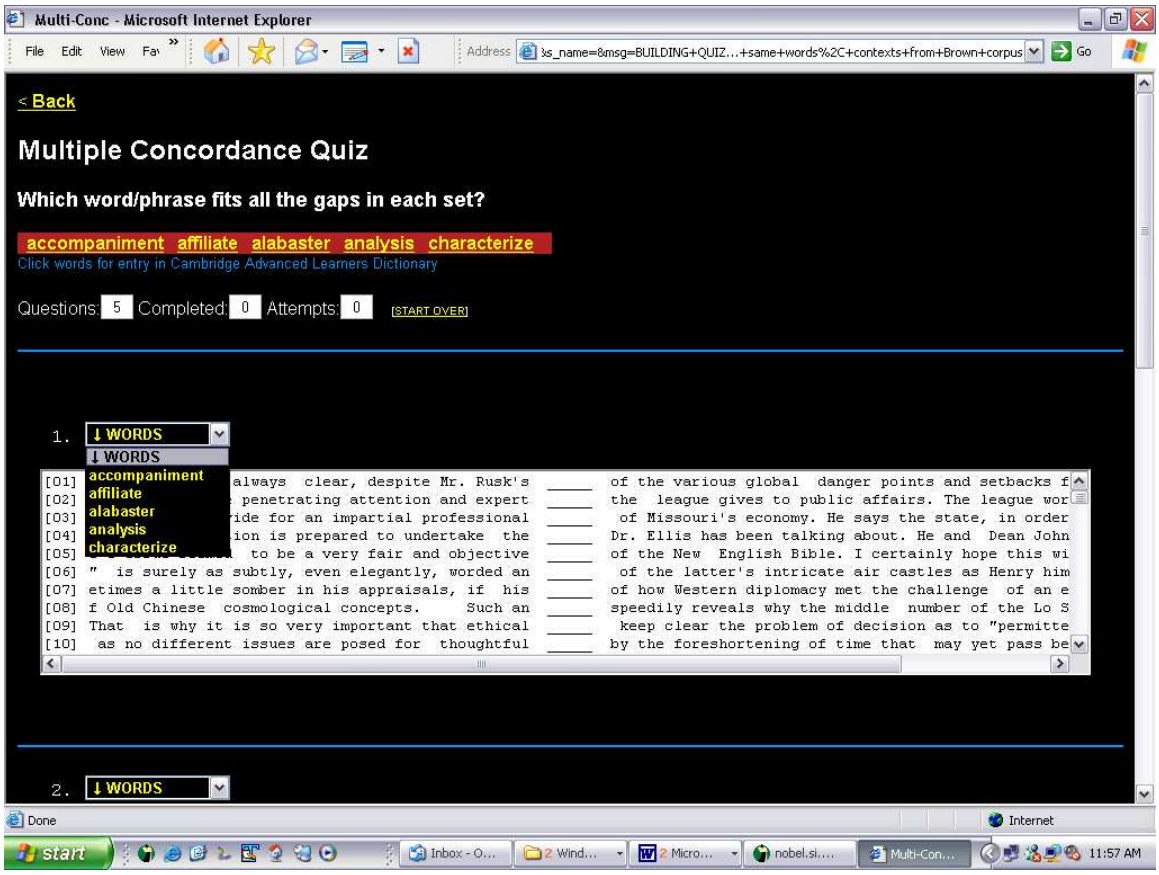

Figure 17: Transfer to novel contexts - revisited

To conclude Part II, it seems clear that properly designed computer programs, properly used, can substantially increase the number of exposure to new words through reading. But is it substantially enough? Both quantification and empirical investigation are waiting to be completed. In the mean time, computer programs can not only increase the number of exposures, but also help teachers and learners do more with the exposures available. This is the topic of the next section.

Part III: Improving the quality of individual exposures

Up to now, this paper has shown a number of ways computers can increase the number of exposures to words. Now we turn to a different dimension -- what computers can do to improve the quality of an individual exposure. When a new word is met, there are two things a learner can do with it if he or she decides to give it some attention. One is to look

it up in a dictionary, and there are many high quality learner dictionaries now available for this purpose. The other is to attempt to infer a meaning from the ongoing context. However, both these strategies present problems. Dictionaries take the reader out of the text, physically and mentally and almost certainly disrupt the flow of reading. Contextual inference (and probably successful dictionary use as well) is reliable only if 95% of the words in the context are known (Laufer, 1992; Nation, 2001), and this is rarely the case for all but the most advanced learners. Reading on a computer may be able to address both these problems.

Problem 1: High-quality dictionaries can improve text readability, but at the same time, they disrupt the flow and possibly the pleasure of reading.

Several publishers of ESL materials -- notably Longman and Cambridge -- have recently invested in well researched and designed learner dictionaries for intermediate and advanced learners. Nonetheless, studies of dictionary use (e.g., Hulstijn, Hollander & Greidanus, 1996) suggest that, however advantageous, even sophisticated learners may believe these dictionaries to be, they will not use them extensively if they believe them to entail an exit from the reading task itself. This, of course, might not happen as much if the resources could be directly integrated into the text that the learner was reading. This ideal has recently become possible as free online versions of these dictionaries are now available.

This is precisely the object of the dictionary option at http://www.lextutor.ca/hypertext. The reader copies a text into a Web form, chooses from a menu of four excellent online dictionaries (including the online versions of the Longman *LDOCE* and the new *Cambridge Advanced Learners'*), and the program wires text and dictionary together such that a click on *any word* in the text produces the relevant definition in a window just beside the text. (It should be noted that this any-word feature depends on the fact that all of these dictionaries are fully lemmatized, or fleshed out as word families, so that clicking on *cats* in the text produces the entry for *cat*, for instance, rather than to a 'Not Found' notice.) The learner working in Figure 18 has connected the Cambridge Advanced Learner's Dictionary to a text on Cell Phones and Driving and run immediately into the unknown word *ban*. A click on the word generates a well thought out definition in roughly one second. Is this significant disruption or not?

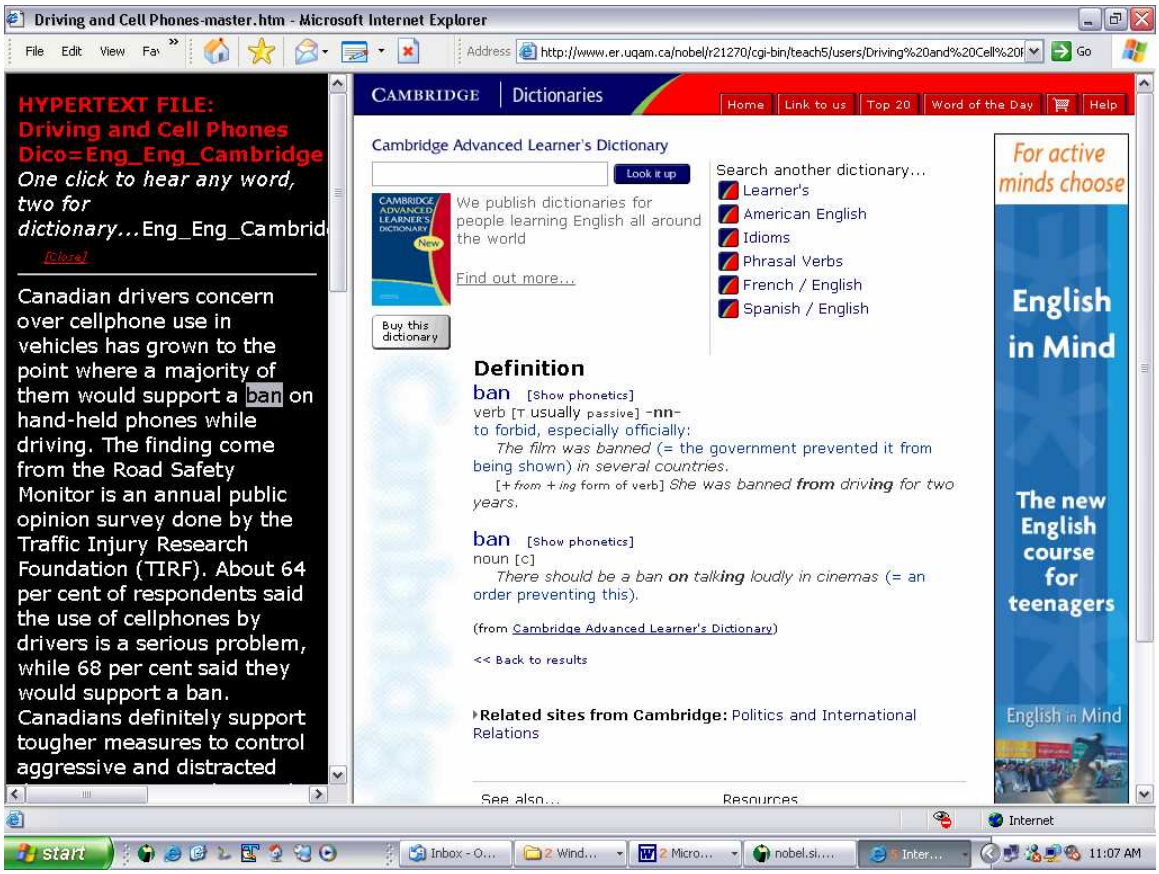

Figure 18: High-quality any-word click-up online definitions

Research: A repeated-readings case study by Cobb, Greaves and Horst (2001) compared two Anglophone learners reading similar-sized extensive texts in their target languages -- one (a German learner) reading a German novella on paper, and the other (a French learner) reading a French version of the *Call of the Wild* page adapted for Guy de Maupassant's novella, *Boule de Suif*. Vocabulary expansion resulting from the readings was used as the measure of their success and learning value. The offline reader simply read his text the required number of times, while the online, resource-assisted reader could access a dictionary and several other resources on a click-on basis. Learning was tracked for several hundred single-occurrence words in both texts (a quantity of test items made possible through the employment of a computer). Vocabulary growth was roughly double for the dictionary-linked, online reading experience, which was perhaps not surprising. More interesting, however, for the present research question, was the fact that time-on-task was no greater for the online reader. In other words, look-ups were not consuming large amounts of reading time and seem to have been an adjunct to, rather than a disruption to, reading, as indicated by the subject's report as well as the time record. Unfortunately, no similar experiment has yet been undertaken for a larger group of learners.

Research – open questions: Our research was merely preliminary. It is still not fully established whether easy look-ups do or do not constitute a significant disruption from

the text. The study waiting to be done is one that would compare online and offline resources for reading comprehension.

Problem 2: Learners learn to read by reading, yet most L2 texts are simply unreadable by most learners.

While text comprehensibility is somewhat linked to learner strategies and topic familiarity, as a general rule, a readable text is one in which 95% of the vocabulary is known to the reader, or, in other words, in which the new-to-known ratio is 1:20. This is the point at which comprehension tests can be passed, new vocabulary reliably inferred and reading is less effortful and more pleasurable (Laufer, 1992; Nation, 2001). But how can such texts be located, or created?

Graded readers are often able to provide opportunities to meet new words in low-density environments, and there can be little doubt that these readers should be in far greater use than they are at present. However, there are two problems with implementing graded readers: they tend toward children's interest levels (adventure stories and the like), and it is difficult to build a collection that caters to a wide enough range of interests -- particularly adult interests. Successful graded reading programs ultimately depend on a growing supply of teacher adapted texts. Yet, how such texts are to be adapted, even just from the lexical point of view, is not obvious. The rest of this paper introduces tools that can help with the job.

Successful grading of texts depends on having some way of defining the lexical levels of both learners and texts, and fortunately it is possible to do this. On one side, vocabulary tests are available that can roughly indicate a learner's vocabulary size in terms of 1000 word-frequency levels (as devised by Nation, 1990; Laufer & Nation, 1999; Schmitt, Schmitt & Clapham, 2001; some of these available on Lextutor). From another side, a computer program is available that analyses texts in terms of this same 1000-level scheme. This program is Nation and Heatley's (1994) *Range*, also adapted for Internet and available through Lextutor under the name Vocabprofile (at www.lextutor.ca/vp). In principle, these two analyses should put texts and learners into communication with each other. Texts can be found, written or adapted to match particular learners' abilities—not easily, of course, but this technology at least makes it possible.

For example, the profile for Chapter 1 of the fictional work *Call of the Wild* shown in Figure 19 indicates that just over 80% of its words come from the first 1000 word families of English, so that a somewhat larger vocabulary than 1000 word families would be necessary for a learner to make any sense of the text. A learner knowing 1000-word families would be facing a new-to-known ratio of about one word in five, not one word in 20. For such a learner, a text with 90% or more at the 1000 level would be more suitable.

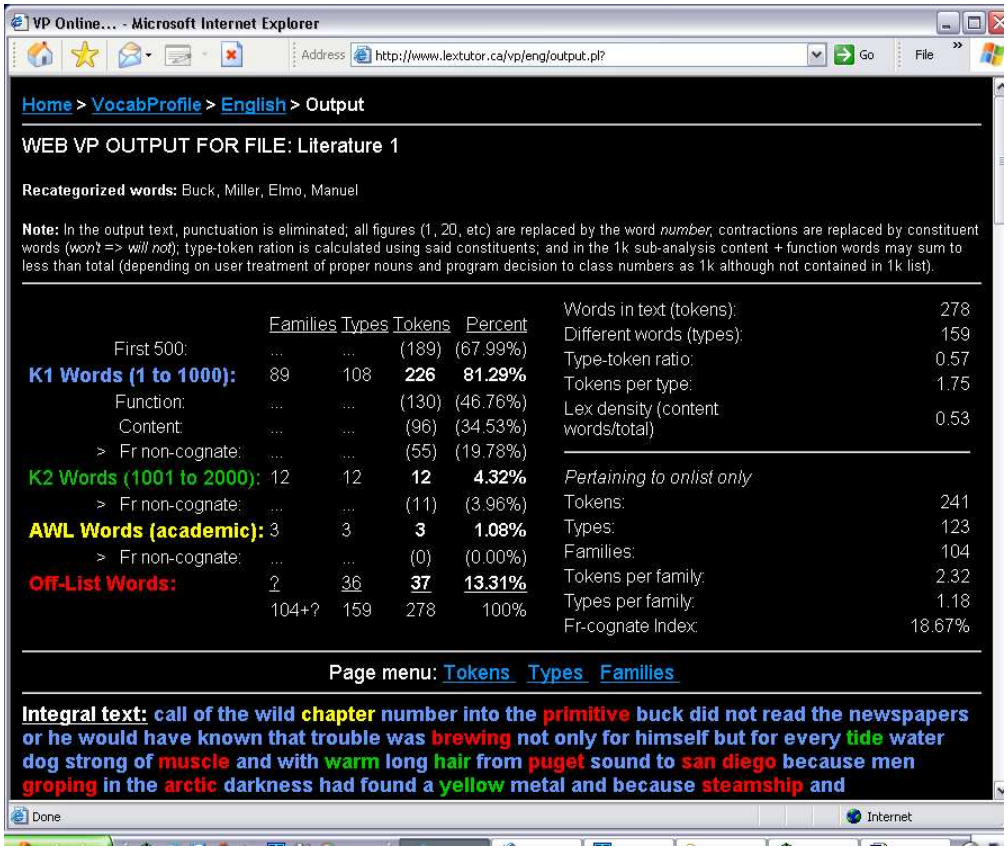

Figure 19: VocabProfile for call of the Wild, Chapter 1

A potential problem with the above analysis, however, is that over the course of an extended text, words are presumably repeated and, thus, have a chance of being learned, so that the new-to-known ratio could be substantially reduced by the end of the story. To what extent does this happen? Are intermediate learners rewarded with reasonable new-to-known ratios if they struggle through the first few chapters of such books as *Call of the Wild* that are written for native speakers?

The answer to this question is provided by another of Lextutor's analytic tools, Text_Lex_Compare, which identifies the new (i.e., different) words in a second text in comparison to those in a first text. This program reads in texts of up to 20 chapters and identifies, counts and lists the new items appearing in each. Further, it automatically links these items to the above-mentioned VocabProfile program for frequency evaluation. This program input is shown in Figure 20, with the first two chapters of the same story in position for a lexical comparison. The output is shown in Figure 21.

Figure 20: Text_Lex_Compare input

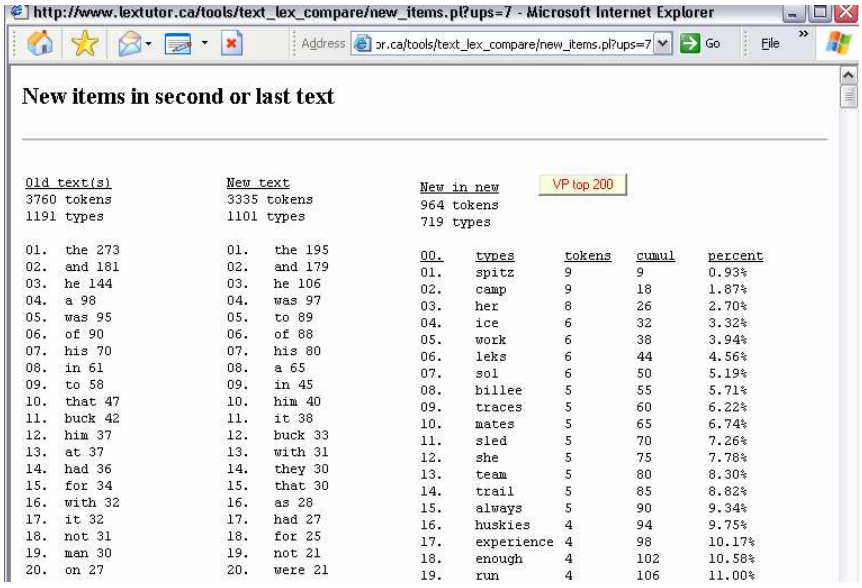

Figure 21: Text_Lex_Compare output

The output shows that there are 719 new different words in the second chapter represented in 964 running words or tokens. *The VP Top 200* button subsequently reveals that 41% of these words are from the most frequent 1000 words of English. In other

words, nearly 60% of these are relatively infrequent for a learner who knows 1000 words. So far, then, this analysis suggests that the lexis of *Call of the Wild* offers a fairly unfriendly lexical ratio for the intermediate learner, but the pattern must be worked out for the rest of the text – made possible by the series of upload inputs in the bottom half of Figure 20. For example, the analysis of Chapter 3 shows how much new lexis it presents with respect to both of the preceding chapters, and so on. The result of this analysis for the entire volume is shown in Table 7.

Chapters	New types	New tokens	Per cent 1k items
2	743	982	37
3	800	983	29
4	351	401	27
5	632	893	36
6	565	733	32
7	633	795	25
Mean	620.7	797.8	31.0

Table 7: New lexis chapter by chapter

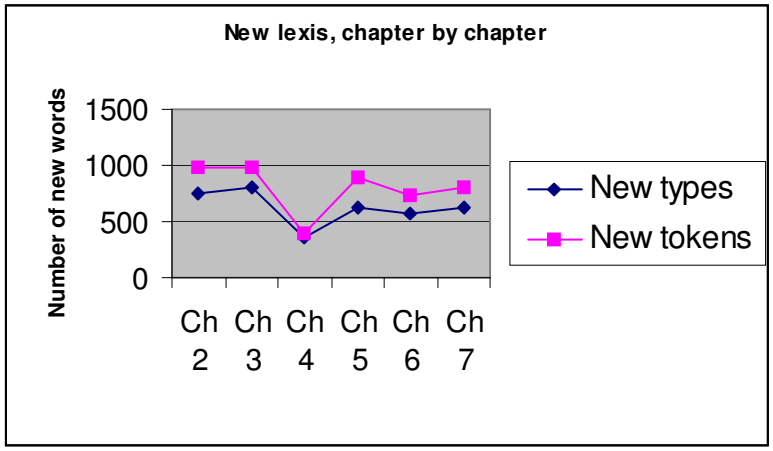

Figure 22: Little reduction in the diet of novel lexis

Table 7 clearly shows that the flow of new lexis never abates in a text designed for native speakers. Indeed, the third highest number of new word types appears in the final chapter. Further, the Vocabprofiles for these new items show them to be mainly post first-2000 items -- potentially difficult items that are fairly rare and will not necessarily repay any investment in learning. But the new-to-known ratio is the biggest problem. With the exception of the third chapter, these chapters are about 3500 words in length, and the number of new word tokens is an average of nearly 800 words. The ratio, in other words, is about eight new words per 35 -- more than one new word in four—rather far from the 1 in 20 mark!

Similar findings are available for other books written with native speakers in mind. Conan Doyle's *Hound of the Baskervilles* can be downloaded from Lextutor's Text_Lex_Compare page for readers to test this assertion for themselves.

While it is a truism that our learners ought to be reading graded texts, with the research findings and technologies now at our disposal, we are in a position to see clearly why this has been proven. Text_Lex_Compare was fed with the seven chapters of the Penguin/Longman graded version of the same *Call of the Wild* with results as shown in Table 8 / Figure 23.

Chapters	Total words	New types	New tokens	Per cent of 1k items	New-known	Per cent	Ratio
Ch 2	876	131	193	79	193/876	0.22	1 to 4.5
Ch 3	1573	116	199	64	199/1573	0.12	1 to 8.5
Ch 4	1272	69	111	56	111/1272	0.08	1 to 12.5
Ch 5	1178	49	112	40	112/1178	0.09	1 to 11.5
Ch 6	1584	63	139	45	139/1584	0.08	1 to 12.5
Ch 7	1838	50	110	53	110/1838	0.059	1 to 19
MEAN	1386.8	79.7	144.0	56.2			

Table 8: results for graded *Call of the Wild*

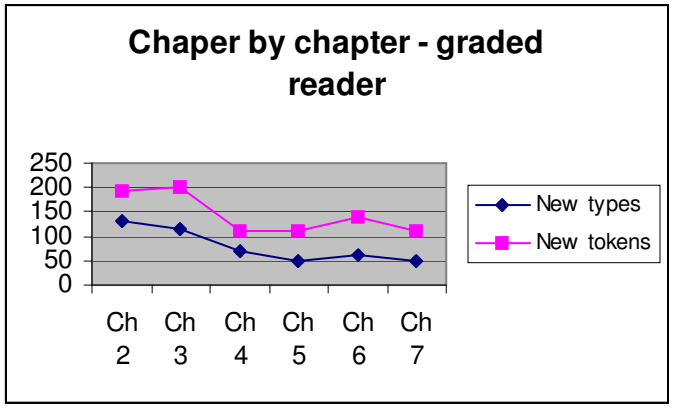

Figure 23: Fewer and declining number of new word types

The adaptors of this text have produced a far more manageable proportion of new lexis, more than half within the first 1000 level (56.2%); moreover, a decreasing amount of it as the novel proceeds -- indicating a good deal of recycling of old items. Most interestingly, the new-to-known ratio has actually come close to the 1-in-20 target in the final chapter (1:19 to be precise, as shown in the final ratio of Table 8). This means that if someone learned all the words that appeared in the previous chapters, then by the final chapter he or she would be reading with a new-word density of just more than one new word in 20. For the other six chapters, of course, the density is higher than that, which I believe is an argument for linking even simplified readers to a relevant selection of the computer-based learning resources described above. And this, in turn, is an argument for providing learners with simplified extensive texts that are machine-readable.

Where would a varied, multilevel library of machine-readable extensive readings come from? Probably not from the big commercial publishers. While companies like Longman have produced an impressive paper collection of graded materials, mainly in the realm of fiction, they have also been quite successful (unlike music publishers) at making sure that teachers and learners pay for anything they get. This is not a criticism of the publishers; text adaptation is hard work, and the publishers are entitled to recoup their investment.

The Internet is lacking in very few types of texts, but one of the few is simplified reading materials for language learners. While preparing this piece, I sent out a plea to the extensive reading community (via the Extensive Reading Pages website) to inform me of any free online sources of extensive readings and learned that there is apparently only one in existence -- an interesting but modest UK site named *Blue Yonder*. This lack of available materials led me to conclude that a complete and useful library of graded readings, particularly, at adult interest levels and including non-fiction texts as well as fiction, with vocabulary level and new word density rates publicised, can probably be produced only by teachers and course designers themselves. This would not be easy, but, of course, the benefits could be shared over the Internet -- perhaps at a dedicated Website. The resource Text_Lex_Compare both shows us why we need to do this, and, along with VocabProfile, provides tools to help with the job.

How would the simplification process work? Assuming a text is in machine-readable format, it can be run through VocabProfile and its potentially difficult or unuseful vocabulary identified (difficult in terms of the intended readership's lexical profile). Decisions would then be made as to whether each word was a proper noun or place name posing no problem (and, hence, could be recategorized as a common item), was a crucial or repeated item requiring a contextual gloss in more basic language or was neither and, therefore, should be written out of the text. When the profile of the text as a whole shows roughly five per cent challenging, yet learnable, vocabulary, the chapters can then be run through Text_Lex_Compare in sequence to determine whether the density is proportional over the course of the text. If not, further modifications are necessary. This is not easy work; anyone who has done it knows why Longman and other publishers guard their simplified readers so jealously. But, with these computational tools, it is feasible work.

Conclusion

I hope I have convinced the reader that the role of the computer within the expanded universe of text can, and should, go well beyond the functions of delivery, distribution and printing. At the beginning of this paper, I proposed that "computer programs, accessing large shared text repositories, have a tremendous potential to both resolve old questions for teachers/course designers and provide new and unique opportunities for large numbers of learners at low cost." Within one domain, extensive reading, and using vocabulary growth as the index of the success of extensive reading, I have shown how corpus analysis can define some of the key problems with growing a lexicon through reading and how the networking of different kinds and forms of texts on the learner's computer screen should be able to solve it.

The key problems of learning through extensive reading are clear. Range analysis shows that post-1000 level words are unlikely to be encountered in natural reading in sufficient numbers for learning to occur. VocabProfile shows that the amount of new vocabulary in natural texts is likely to be severely at odds with both the lexical level and learning capacity of intermediate learners. Text_Lex_Compare further shows that the rate of new-word introduction in a text designed for native speakers is far more than these learners are able to cope with. And, yet, these same tools can also be employed positively, to help with the adaptation of texts that learners can read and learn from.

The long-term goal is to build a shared, free online, universal library of graded reading materials. The short term goal is to get available texts online and help learners use the tools described above, which can proliferate encounters, keep track of encounters, multicontextualize meanings and provide links to top quality learning tools—all with minimal reading disruption. Only local teachers and course designers can accomplish any significant part of this in any coherent way. Ironically, Lextutor's records and user correspondence show that the main users of the Websites learning tools at present are individual learners. In other words, the market is there.

Krashen (1989) remarked in a deservedly famous paper on vocabulary growth from reading that a number of books can be purchased for the price of one computer, implying that the books were the wiser choice. In 2005, books and computers seem less a choice than a partnership.

References

Blue Yonder website for extensive reading. [Online]. Accessed 2005 April 20, at http://gradedreading.pwp.blueyonder.co.uk .

Cobb, T. (1997a). Is there any measurable learning from hands-on concordancing? *System*, 25(3), 301-31.

Cobb, T. (1997b). *From concord to lexicon: Development and test of a corpus-based lexical tutor.* Concordia University: Unpublished PhD dissertation. [Available at http://www.er.uqam.ca/nobel/r21270/webthesis/Thesis0.html.]

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development, 47* (3), 15-33.

Cobb, T., Greaves, C., & Horst, M. (2001). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la didactique des langues secondes* (pp. 133-153). Montréal: Éditions logique.

Cobb, T. & Stevens, V. (1996). A principled consideration of computers and reading in a second language. In M. Pennington (Ed.), *The power of CALL* (pp. 115-136). Houston: Athelstan.

Extensive Reading Pages website. [Online: http://www.extensivereading.net/.]

Heatley, A. and Nation, P. (1994). *Range*. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Hirsch, D, & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language*, 8(2), 689-696.

Horst, M. (2000). *Text encounters of the frequent kind: Learning L2 vocabulary from reading*. University of Wales (UK), Swansea: Unpublished PhD dissertation.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond A Clockwork Orange: Acquiring second-language vocabulary through reading. *Reading in a Foreign Language,* 11, 207-223.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding Academic Vocabulary with a Collaborative On-line Database. *Language Learning & Technology,* 9,(2), 90-110.

Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review,* 56,(2), 308-328.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal,* 80*,* 327-339.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal,* 73*,* 440-464.

Krashen, S. (2003). *Explorations in language acquisition and use: The Taipei lectures.* Portsmouth NH: Heinemann.

Kucera, H., & Francis, W. (1979). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers* (Revised and amplified from 1967 version). Providence, RI: Brown University Press.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P.J. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: Macmillan.

Laufer, B. & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing* 16(1), 33-51.

Mondria, J-A., & Wit-De Boer, M. (1991). Guessability and the retention of words in a foreign language. *Applied Linguistics,* 12 (3), 249-263.

Nation, P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Simpson, R., Briggs, S., Ovens, J. & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing,* 18(1), 55-89.

Stanovich, K.E., & Cunningham, A.E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition,* 20*,* 51-68.

Wesche, M., & Paribakht, S. (1996). Assessing vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review*, 53(1), 13-40.

Zahar, R., Cobb, T. & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, 57(4), 541-572.

Learning Academic Vocabulary with On-line Concordancing

Marlise Horst

How does the use of technological tools to study a second language affect learning? Can technology make a distinct and measurable difference? These are the ambitious questions that the research reported in this chapter attempts to answer. The project is ambitious because learners obviously acquire new language knowledge in many ways and it is difficult to disambiguate the effects of studying with the aid of computers from the effects of such other factors as skilled classroom teaching, exposure to language outside of English as a Second Language (ESL) class, or study using more traditional techniques, to name just a few. The investigation reported here attempted to minimize the effects of the many possible variables by focusing rather narrowly on a single technology tool and a single strand of language. The setting for the study was an experimental vocabulary course offered to learners of English at a university in Montreal in the fall of 2004. The students used a number of computer resources designed by Cobb (see Cobb, this volume, for a description) to study many words from a variety of sources. However, the focus of the study reported here is students' use of one of the tools, an on-line concordancer, as they studied a specific set of words, Coxhead's *Academic Word List* (2000).
The research set out to determine the extent to which students used concordancing in their study of *Academic Word List* (AWL) items and the extent to which vocabulary knowledge gains they achieved can be linked to use of this on-line study tool. Findings of the investigation are reported below. But first, reasons for focusing the investigation on concordancing the AWL are discussed.

Why concordancing?

Concordancing software searches a large corpus of texts for instances of a target word and displays lines of contextualized uses in a convenient format. Twelve lines from a concordance search of the word *construction* (using tools available at www.lextutor.ca) are shown below by way of illustration.

1 go, when my houseboat home/office was under construction, a struggling company called International S
2 nforcing cement. Kajima, the giant Japanese construction company, has been working with Otani to creat
3 reductionist conception of form in space as construction. ' Comparative arts/exhibits review. Like hi
4 Valley. Along with the estimated $5 billion construction cost, the plant would cost more than $300 mi
5 ntury mishaps will have to be factored into construction costs and covered by adequate insurance, an
6 Grand Prix sputtered to a start Saturday as construction crews rushed to finish fencing the track, o
7 's economy, starting farms, businesses and construction crews. Now it is playing a larger political r
8 organized quickly. A plan to finance school construction in Prince George's County by taxing new deve
9 transportation of building materials. When construction is completed at one site, the mobile product
10 res fewer workers, and has an uncomplicated construction. It may sound crazy, but technology at times
11 lso includes a narrative description of the construction of an AFCARS Annual Database. The organiz
12 en variables. Let's first look at the basic construction of graphs. A graph is a visual representation

Figure 1: Concordance of the word *construction*

In fact, the concordancer located 94 context lines but a great deal of useful information about this word is readily apparent in the first 12 lines that are shown in Figure 1. In terms of meaning, it is clear that *construction* usually refers to the building of outdoor

structures such as houseboats (line 1), plants (line 4), fences (line 6) and schools (line 8), but it can also be extended to other projects such as creating databases (line 11) and graphs (line 12). In terms of the word's formal properties, *construction* is clearly a noun that can be used in noun-noun combinations such as *construction cost* (lines 4 and 5) and *construction crew* (lines 6 and 7). There is also the collocation pattern '*the construction of + noun*', a pattern that appears in line lines 11 and 12 and recurred many more times in the 82 other concordance lines that followed these 12. A concordance search of a slightly different type can also reveal a great deal of additional information about derived and such inflected forms of the root verb *construct* as *constructed, deconstruction, constructive*, etc.

A concordancer linked to a large corpus of authentic texts is clearly a rich resource to put in the hands of language learners. Concordance assisted vocabulary learning is also a good test of computer assisted language learning (CALL) since it requires a computer to implement it. Such activities that typically appear on ESL vocabulary websites as matching words to definitions or filling in gapped sentences might be implemented just as easily and effectively without computers (e.g by using such traditional media as flashcards or worksheets). But this is clearly not the case with concordancing. While classroom teachers may add a sentence example or two when they present the meaning of a word like *construction,* they cannot easily match the concordancer's ability to search millions of words of text for instances of a target word, locate many examples, and display them to learners in a matter of seconds. Since examining concordances to identify patterns of use or infer meaning requires effort on the part of the learner, there is good reason to think that the activity will foster retention of new form-meaning associations and patterns of use (Laufer & Hulstijn, 2001).

A number of studies have explored the link between concordancing and the acquisition of new second-language word meanings (Cobb, 1997; Cobb 1999, Cobb and Horst, 2001, Horst, Cobb & Nicolae, 2005). An aspect not explored in these studies is the extent to which learners who use concordances also acquire knowledge of the collocational and morphological aspects of new words – aspects that concordance searches can make salient in ways that are not easily replicated by dictionaries or teachers. The investigation reported here took on this added dimension by including an essay writing task as one of the experimental measures. This allowed the examination of both the semantic and formal accuracy of words students used in their essays before and after participating in the experimental course.

Why the AWL?

Learning a second language is a vast enterprise and the sheer enormity of the undertaking presents a perennial quandary for ESL teachers and course designers: Given that only a limited number of the multitudes of words and structures can be presented in a single course, which ones are the most important to focus on? Fortunately, in the case of the vocabulary needed for academic study in English, a clear answer has been provided by Coxhead (2000) in the form of the *Academic Word List* (AWL). This 570-word list is the result of an analysis of a large corpus of passages from university textbooks (3.5 million

words). The goal was to identify words that occurred both frequently and consistently across domains of study. Examples of words from this useful list are *apparent*, *element* and *reaction* – words that can feature as easily in a description of the movements of atoms as in a critical examination of rococo painting.

The value of knowing the AWL for university ESL learners is demonstrated in the following set of passages taken from a business and finance textbook (Buckle, Kim & Hall, 1994, cited in Coxhead, 2000). In Version 1 below, I have blanked out words that are not on West's (1953) list of the 1000 most frequent English word families. Version 1 is intended to simulate the experience of a low proficiency learner of English who is faced with the task of reading this text knowing only 1000 frequent English word families (proper names and numbers appear because they are assumed to be lexically transparent). The blanks represent less frequent words that this learner would not be able to understand. It is obvious that this reader would have a great deal of difficulty processing the text.

> Version 1 (0-1000 most frequent words shown)
> Dating the turning points and _____ of business _____ has long been associated with the _____ of _____ _____ _____ _____ and their associated leading, _____ and _____ _____. This was along lines _____ developed by Burns and Mitchell (1996), and _____ by _____ at the National Bureau of _____ _____ (NBER), e.g. Klein (1990). More recently, _____ the turning points and _____ of business _____ has been an important _____ of two further _____ of business _____ _____: the _____ of _____ and associated _____ business _____ _____ and the _____ of the time _____ characteristics of business.

If we assume that our hypothetical ESL reader has acquired knowledge of the meanings of the next 1000 most frequent words of English, his/her position is somewhat improved as can be seen in Version 2 below. Here I have added the words from West's (1953) list of the 2000 most frequent words with less frequent words still appearing as blanks. Yet it is clear that even with this increased vocabulary knowledge, the text will remain difficult for the reader to process with several dictionary look-ups required in every line.

> Version 2 (0- 2000 most frequent words shown)
> Dating the turning points and _____ of business _____ has long been associated with the _____ of _____ _____ _____ _____ and their associated leading, _____ and _____ _____. This was along lines originally developed by Burns and Mitchell (1996), and _____ by _____ at the National Bureau of _____ _____ (NBER), e.g. Klein (1990). More recently, _____ the turning points and _____ of business _____ has been an important _____ of two further _____ of business _____ _____: the _____ of _____ and associated _____ business _____ models and the _____ of the time _____ characteristics of business _____

Version 3 of the text simulates the experience of an intermediate university ESL learner with knowledge of 2000 frequent English words in place, and, importantly, our hypothetical learner has also had the opportunity to learn the items on the AWL. The impact of knowing the AWL words is dramatic: Even though two words are still blanked out, the text is suddenly comprehensible and it becomes possible to infer the meaning of at least one of the 'missing' words. It may come as no surprise to many readers that in the original text, the word represented by a blank in "….their associated leading, coincident and _____ indicators" (see Version 3) turns out to be *lagging*. (The other blanked word was *reference*.)

> Version 3 (showing 0- 2000 + AWL)
> Dating the turning points and duration of business cycles has long been associated with the construction of aggregate _____ cycle indexes and their associated leading, coincident and _____ indicators. This was along lines originally developed by Burns and Mitchell (1996), and subsequently by colleagues at the National Bureau of Economic Research (NBER), e.g. Klein (1990). More recently, identifying the turning points and duration of business cycles has been an important aspect of two further areas of business cycle research: the evaluation of theoretical and associated empirical business cycle models and the analysis of the time varying characteristics of business cycles.

The potential of the AWL to make a difference in university ESL learners' ability to read academic texts (as demonstrated in the text sequence above) has been a compelling rationale for giving the AWL a prominent place in the syllabus of earlier versions of the experimental vocabulary course described in this chapter. But a consistent theme in feedback from students in the earlier courses was a desire to move beyond comprehension to active use of the vocabulary in writing. This request for more emphasis on production came just as I became aware of the arrival of a new ESL textbook that specifically targets the AWL entitled *Focus on Vocabulary: Mastering the Academic Word List* (Schmitt & Schmitt, 2005). The authors allowed a pre-publication version of their book to be used as the core text in the experimental vocabulary course in the fall 2004 course. Their materials present AWL words in authentic academic texts supported by a variety of exercises, many of which require students to use AWL words productively.

A variety of technology tools were available to support students' study of the AWL words (see Cobb, this volume, for more on these resources). One of these was an on-line Word Bank; every week each student contributed words to this study resource along with definitions and example sentences as part of the required coursework. A sample of student entries is shown in Figure 2. The Word Bank was configured to allow students to select particular items for study (note checked boxes for *demonstrate*, *denote*, and *derive* in Figure 2) and build a personalized quiz of problem items. An example of a partially completed quiz that was created in this way is shown in Figure 3.

119	☐	**deduction**	Astronomists make **deduction**s about the weather on different planets based on photos of atmospheric patterns that are sent by aerospace ships.	Noun	something that is inferred by carefully examining the facts
120	☑	**demonstrate**	The salesman **demonstrate**d this new washing-machine.	Verb	To present by examples, explain and illustrate somthing.
121	☑	**denote**	In England the colour purple **denote**s fortune so it is called the royal`s color.	Verb	To mark, indicate
122	☐	**derivative**	The reason for getting **derivative** results was a coincidence.	Adj	copied or modified from another
123	☑	**derive**	I usually **derive** knowledge from books.	Verb	To obtain or receive from a source.

Figure 2: A sample of student-produced AWL entries in the on-line Word Bank

CHECK * Quiz - words - 21 Apr 05, 13:52

	New word	Example	Part of speech	Definition
1		The salesman demonstrate d this new washing-machine.	Verb	To present by examples, explain and illustrate somthing.
2		In England the colour purple s fortune so it is called the royal`s color.	Verb	To mark, indicate
3		I usually knowledge from books.	Verb	To obtain or receive from a source.

Figure 3: Individualized quiz generated by selecting *demonstrate*, *denote* and *derive* in the Word Bank

Students were strongly encouraged to continue on to a 'Tougher Quiz' option that was linked to the Word Bank. This Multi-Concordance option presents 10-line clusters of gapped concordance lines for each of the selected items and a drop down menu with answer options. Two clusters are shown in Figure 4. Since class quizzes used this question format, it was expected that students would be motivated to try this exercise on their own. Students were also shown how to do their own independent concordance searches like the *construction* example above using the concordancer available at the class website. This was presented as a way to test meaning of unfamiliar words ("Does my meaning hypothesis fit these examples?") and as a way of exploring formal aspects ("What do these examples tell me about part of speech? Possible forms of this word? Its collocations?"). The class website also offered links to on-line dictionaries.

[01] w anyone seeking a position in public life could _____ such poor judgment and bad taste. "Such a vicious sta
[02] e of this danger; therefore we need guardians to _____ against the ghastly stupidity of nuclear weapons and jolt
[03] is work is no favorite of mine. I am prepared to _____ at an ytime that it represents the spirit of Imperial Rus
[04] ung men from the audience will take the floor to _____ their own amateur graces. Except for the odd uptown sex m
[05] nd British Broadcasting Corporations continue to _____ . Unfortunately, "Poised for Violence" was not the happies
[06] ocosm, one might reasonably have expected him to _____ precise knowledge not only of techniques but of mores an
[07] hrens of the Rockefeller Institute, took over to _____ the chemical difference between vegetable and animal fats
[08] il treatment alone that accomplished this. To _____ the soil of his vegetable gardens as it is today, Mr. Cla
[09] loved to dance in the nude, something she was to _____ time and again. She developed another quaint habit. Ev
[10] not an edifying Trial have made every effort to _____ this once and for all by showing how representative types

[01] the total work of the university. Religious who _____ their own sense of purpose through identification with the rel
[02] ether partial to music or no, you can't help but _____ joy from this picture if you have a sense of humor and a heart
[03] to deny these purists the obvious pleasure they _____ from all this, and to give fair warning where warning is due,
[04] e in. Apart from the categorical imperative they _____ from the metaphysics of the orgasm, the only affirmation they
[05] claration of Independence says that "governments _____ their just powers from the consent of the governed". The phras
[06] bilities in history. These differences, in turn, _____ from prior differences concerning the friendly or hostile char
[07] come up in aesthetic discussion, seem partly to _____ their import from the "renewal" of purpose and a "refreshed" s
[08] The general theory of resonance shifts is used to _____ a general expression for the second moment **f of a polycryst
[09] uld make it clear that countries which choose to _____ marginal advantages from the cold war or to exploit their pote
[10] array indispensible to that best ordering. To _____ Utopian communism from the Jerusalem Christian community of th

Figure 4: Two clusters in gapped concordance format, a "tougher quiz" (answer options: *demonstrate*, *denote*, *derive*)

The extent to which students used the various resources and the concordance-based techniques in particular are detailed in the next section. First, word-learning outcomes and levels of resource use in the entire group are examined. Then I take a closer look at four students and explore links between their use of the concordancing options and gains in semantic and formal accuracy in their use of AWL words in written productions. The research questions are as follows:

1. To what extent did students in the course increase their knowledge of AWL words?
2. To what extent did students make use of the on-line resources?
3. What do learner profiles reveal about the learning effects of study using on-line concordancing?

Methodology

Participants

There were 17 participants in the 13-week course that was investigated in this quasi-experimental study; the course took place twice a week for a total of three hours per week. The students were of intermediate-level proficiency; all had taken the university's placement test and scored just under the minimum criterion for placement in the university's regular credit ESL program. Most had opted for the experimental course with the goal of scoring higher when they took the placement test again at the end of the semester. First languages represented in the group were Arabic, Chinese (Mandarin), Farsi, Hungarian, Japanese, Spanish and Vietnamese. Study interests included accounting, electrical engineering, management, psychology, sociology and theatre. Four students -- two who used the concordancing options often and two who used them rarely

or not at all -- are described in more detail in a later section. A graduate student in the university's Applied Linguistics program with experience in teaching ESL and a special interest in computer-assisted language learning was recruited to teach the experimental course.

Measures

At the outset of the course, students' knowledge of AWL items was assessed using the Vocabulary Levels Test (Schmitt, 2000). This measure assesses ability to recognize simply worded definitions in a clustered multiple-choice format. A sample cluster is shown in Figure 5.

1. consent	
2. enforcement	____ total
3. investigation	____ agreement or permission
4. parameter	____ trying to find information about something
5. sum	
6. trend	

Figure 5: Sample test items from the AWL section of the Vocabulary Levels Test (Schmitt, 2000)

A second measure assessed productive use of AWL words. Students were given 30 minutes to respond in writing to the following prompt:

> Choose a topic & write a short paragraph on one of the following:
>
> The importance of saving the environment
> The importance of communication between old and young people
> How life was different 300 years ago
>
> Use 5-10 words from the list below:

contemporary	diverse	evidence	item
consequences	element	evolve	global
manipulate	neutral	contrast	estimate
encounter	source	further	decline
symbolize	interact	generation	eventually
capacity	environment	transform	complex

Both of these measures were administered again at the end of the course. In addition to the research measures, students also took five 20-minute in-class AWL quizzes and a final exam that covered AWL vocabulary and other words that had been entered in the class Word Bank. Students also responded to an end-of-course questionnaire that explored attitudes to the experimental materials and the extent to which they had made use of the on-line tools. The three questions that assessed use of concordancing activities were as follows:

How often did you use the on-line concordance to look at examples?

How often did you use the on-line concordance to get collocational information about words (e.g. consist + of)

How often did you make a gapped concordance exercise?

For each question, there were four answer options:

A. at least once for every chapter (i.e. very often)

B. at least once for every unit (i.e. fairly often)

C. rarely

D. never

Analyses

To determine whether gains in knowledge of AWL words had been achieved during the experimental course, I compared pre- and post-course mean percentages of correctly identified words on Vocabulary Levels Test for the14 students who were present for both sittings of the test. A *t*-test was used to determine whether differences in mean scores on this recognition measure were statistically significant.

Changes in productive use of AWL words were assessed by examining the semantic and formal accuracy of AWL words in writing samples produced at the beginning and the end of the course by four learners (procedure is detailed below). Two of the participants were identified by their teacher (and by survey results) as being enthusiastic concordancers while the other two were clearly not predisposed to use this study method. The pre- and post-course essays of these four learners (eight essays in total) were transcribed and corrected for spelling[1] so that the AWL words that occurred in them could be identified electronically. The software used for this purpose was *VocabProfile*, a program available at www.lextutor.ca, that tags each word of a text as belonging to the list of the 1000 most frequent word families, the 1001-2000 most frequent list, the AWL, or as being 'off-list' (i.e. not on any of the frequency lists). Of interest in this study were the numbers of AWL words used by the participants, and the extent to which they were used accurately – before and after the experimental course. Two research assistants who teach ESL and are native speakers of English were asked to read the essays and rate the underlined AWL words. For each word, they were asked to make two judgments: 1) semantic, i.e., whether the meaning fit the sentence and overall text, and 2) formal, i.e., whether the usage was accurate. The main examples of possible usage problems noted in the instructions to the raters were morphological and collocation errors. The semantic and formal uses were judged as either accurate or inaccurate; no partial values were assigned. Before they rated the essays from the experimental course, the raters assessed three training essays (produced by ESL learners of a similar proficiency level) and reached agreement on their judgments.

In the excerpt from an essay that appears below, both raters agreed that the participant's use of *environment*, *issue* and *revolution* demonstrated accurate semantic knowledge of these words, but both agreed that there were formal problems with the participant's use of *issue* (pluralization) and *evidence* (collocation).

> There is no doubt that saving the environment is one of the most important issue specially after the industry revolution which began early 20th century…. The evidence at this is ozone hole as an example.

Not all of the decisions were as straightforward as in these instances; nonetheless, an acceptable degree of inter-rater agreement was reached – 89.92% in the case of the semantic judgments and 82.35% for the formal judgments. Percentages of words used with semantic and formal accuracy were then calculated for each of the eight texts. The percentages were calculated by tallying the number of items judged accurate (by one or both raters[2]) and dividing this number by the total number of AWL items used by the writer. For instance, in the case in which the writer used a total of 12 AWL words and the raters judged that nine of these had been used accurately in terms of meaning, the semantic accuracy percentage was 75% (9 ÷ 12 = .75).

Responses to the end-of-course questionnaire that assessed attitudes and levels of use of the various on-line learning tools were analyzed by assigning the response options number values ranging from 0 to 3. For instance, in the case of a question like "How often did you use the on-line concordance to get collocational information about words?," response options were assigned values as follows:

> Never = 0
>
> Rarely = 1
>
> At least once for every unit (i.e. fairly often) = 2
>
> At least once for every chapter (i.e. very often) = 3

This made it possible to calculate group means for each question on the survey.

Results

To what extent did students increase their knowledge of AWL words?

According to scores on the Vocabulary Levels Test, students came to the course with a fairly high level of recognition knowledge of the AWL words. As Table 1 shows, the pre-course mean on this test that requires students to select correct definitions for 30 AWL items was 80% (SD = 12.60). Yet, there was clearly room for growth, and gains did indeed occur. All but one of the 14 students for whom pre- and post-course scores are available scored higher at the end of the course; the post-test mean was 92% (SD = 8.46). The 12% gain is statistically significant ($t = 5.05$, $p < .05$).

	Pre-test %	Post-test %
Mean	80.43	92.43
SD	12.60	8.46

Table 1: Pre- and post-course means for AWL section of Vocabulary Levels Test (n = 14)

To what extent did students make use of the on-line resources?

Survey questions addressed students' use of concordancing for three possible purposes: to infer meanings from sentence examples, to examine collocations, and to make gapped concordance practice exercises. Students were also asked to report their use of the electronic on-line dictionary. Results are shown in Figure 6. The mean rating of 2.24 (SD = .97) speaks to the popularity of on-line dictionaries with ESL learners. The rating is well over 2, which corresponds to fairly often or "at least once for every unit". The slightly lower mean ratings for the concordancing resources approach the 2 (fairly often) level, with inferring meaning from examples proving to be slightly more popular than making a gapped concordance exercise or getting collocational information. A one-way ANOVA for a matched sample was used to determine whether there were significant differences in the data ($df = 3$, $F = 6.76$, $p < .05$). No significant differences were found for the three concordancing activities, but post hoc Tukey tests indicated significant differences between the mean for dictionary use and the two less used concordancing options (making gapped quizzes and gathering collocational information). It is worth noting that the format of the survey probably under-represents the popularity of concordancing, since responses are divided over three uses of this tool rather than loaded on a single factor.

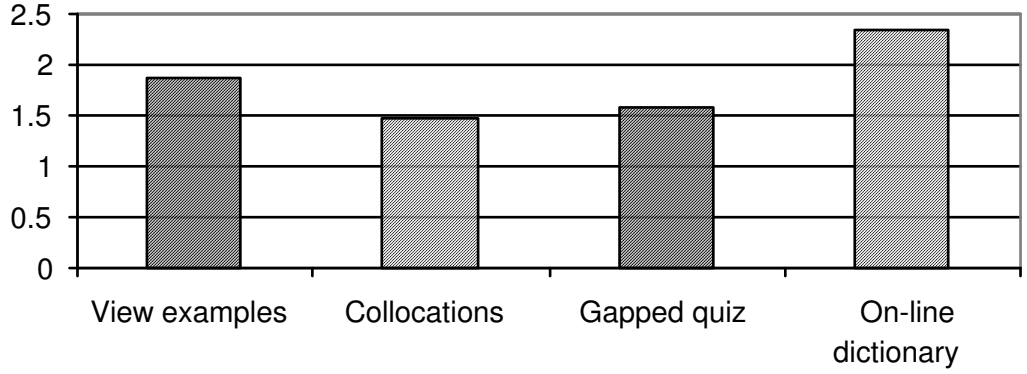

Figure 6: Mean use-level ratings for on-line concordancing and dictionary use (0 = never, 3 = very often)

In sum, the group means for the survey responses indicate that students made regular use of the electronic resources on offer, though it is also clear from sizable standard deviations that use levels varied considerably, with some students using the tools a great

deal and others using them less often. These findings prompted the detailed comparison of users and non-users, which is reported below. But before turning to the four learner profiles, another survey result is worthy of note. In response to the question, "In your opinion, how useful were the readings and vocabulary exercises you did for homework?", the group mean amounted to 2.35 (SD = .70). The 2.35 figure falls between the values of 2 (very useful) and 3 (extremely useful); one of the students chose the "not at all useful" option. Thus it is clear that regardless of their interest in the computerized aspects of the course, the students valued the course materials and the focus on academic vocabulary a great deal.

What do learner profiles reveal?

In this section, the progress of four learners will be examined in detail. The first two, Jamal and Sadaf (not their real names), were identified by their teacher as enthusiastic concordancers. She noted that they were excited about the computer tools made available for use in the course and that they often brought printouts of gapped concordancing exercises they had made in class. Their motivation to use the computer resources is also reflected in their survey responses. Then the progress of two other students is reported. They are Yuan and Hussein, both of whom can be termed non-users on the basis of teacher observations and their survey responses. All four were identified by the instructor as representing middle range performance in the group. That is, they ranked neither very high nor very low in terms of overall class performance on quizzes and on other class work.

User 1: Jamal

Jamal is an Arabic speaker majoring in finance with a fairly high average on class vocabulary quizzes (88%). He reported that he used the gapped concordance to study AWL words frequently, i.e. "more than once a week," and also used the on-line concordancer to look at examples often. The texts that he produced on the pre- and post-course measures appear below with AWL words underlined. Although the later text still contains a number of errors, the increase in number and variety of AWL words used is striking; the writer is clearly able to make a more detailed and fluent argument. Also noteworthy is the evidence of development in knowledge of word forms. The word *environment* recurs several times in both texts, but in the later one, the writer also uses the adjective form *environmental*.

Pre-course text

> There is no doubt that saving the environment is one of the most important issue specially after the industry revolution which began early 20th century.
> Different elements of chemicals which produced from different kinds of industries affect the neutral sources of the earth like water, oceans. etc. The result is the environment will be less pure. The evidence at this is ozone hole as an example.

Post-course text

> Saving environment should be one of the most important priorities for all countries during this century. Evenso the seven giants countries whom are the biggest manufactural countries all over the world, took many procedures to control the different elements which directly affect the environment for instance to restrain earth temperatures from going higher, which called "heat retention phenomena". Other countries specially third world countries did not legislate rigid law clauses to stop environmental pollutions. The United Nations should prohibit all kind of environmental violations through helping the countries to understand the notion "safe environment" and to allocate subsidies for poor countries to virtually teach their nations how to "save environment."

	Pre-course	Post-course
Vocabulary Levels Test (%)	83	90.00
Text: No. of AWL items	12	23
Text: Meaning accuracy (%)	75	86.96
Text: Formal accuracy (%)	83.33	86.96

Table 2: Profile of scores and AWL use for Jamal (user)

Jamal's performance is summarized in Table 2. The first row shows that his score on the multiple-choice test of ability to recognize definitions of AWL words increased from 83% to 90%. Changes in his ability to use them productively in an original text – a much more difficult criterion – are reflected in the next three rows. The most dramatic change appears in the second row where the figures indicate that the number of AWL words Jamal used in his writing increased from 12 to 23 – almost double. The third row shows percentages of words one or both of the native speaker raters judged to have been used with semantic (i.e., meaning) accuracy, while percentages for formal accuracy appear in row 4. It is evident that Jamal's ability to use the words accurately increased; the gain in the semantic aspect is especially impressive. [3]

User 2: Sadaf

Sadaf is a student of electrical engineering and a Farsi speaker. Her average on class quizzes was 82%, somewhat lower than Jamal's. Noteworthy in her reported computer resource use is her attention to use words in context. She rated her use of the concordance "to look at examples" as very high -- one of only two students who reported this level for this purpose. She also reported often making gapped concordance quizzes and consulting the concordance for collocational information often. Her attention to contextual uses is also evident in her response to an open-ended question on the survey that asked students to comment on other study strategies. Here Sadaf notes explicitly that she looks "at other sentences/examples in the on-line dictionary."

	Pre-course	Post-course
Vocabulary Levels Test (%)	70.00	100.00
Text: No. of AWL items	12	21
Text: Meaning accuracy (%)	75.00	71.43
Text: Formal accuracy (%)	91.67	100.00

Table 3: Profile of scores and AWL use for Sadaf (user)

This attention to use of words in context may explain the sizable increase in the formal accuracy of Sadaf's writing as indicated in Table 3.The bottom row of the table shows that post-course uses of AWL forms were rated 100% accurate. Like Jamal, Sadaf also incorporated many more AWL words into her writing by the end of the course - from 12 in the pre-course text to 21. However, it is interesting that Sadaf's levels of semantic accuracy for the production task remain relatively unchanged over time, even though her ability to recognize correct meanings on the Vocabulary Levels Test clearly increased a great deal (from 70% to 100%). It appears that Sadaf's attention to context raises her awareness of both formal and semantic aspects of words but that semantically accurate production lags behind recognition.

Non-User 1: Yuan

Yuan, a Mandarin-speaking student of mechanical engineering stands out as a non-user of the concordancing resources in that he is the only participant to report never having used any of them. In response to the open-ended question about his word study strategies, he wrote "no other way except cram, study hard, read, learn by heart everyday."

	Pre-course	Post-course
Vocabulary Levels Test (%)	97.00	93.00
Text: No. of AWL items	14	9
Text: Meaning accuracy (%)	91.67	100
Text: Formal accuracy (%)	71.43	66.67

Table 4: Profile of scores and AWL use for Yuan (non-user)

He is an interesting case in that he came to the course with a high level of ability to recognize definitions of AWL words, as indicated by his pre-course score of 97% on the Vocabulary Levels Test. This knowledge of word meanings appears to have served him well in terms of semantic accuracy in his writing. His level was already high at the beginning of the course (91.67%) and perhaps rote study of more definitions accounts for the even higher semantic accuracy rating of his post-course sample (100%). But the most interesting aspect of the analyses shown in Table 4 is his comparatively low level of formal accuracy both before and after the course. This student who writes such segments as "The evidences can be picked up easily," and "Contrast to any other countries, …" seems a clear case of a learner who might gain a great deal from study with a tool such as the gapped concordance maker that highlights collocations and other form-related aspects.[4]

Non-User 2: Hussein

Hussein is a student of electrical engineering and a speaker of Arabic. The three options for concordancing – making a gapped self-quiz, looking at meanings in context, or investigating collocations – were all reported as having been used rarely. His mean scores on class tests (63%) suggest that his proficiency in English is substantially lower than that of the other students we have considered so far[5] His writing both before and after taking the vocabulary course is characterized by rather simple sentences that display a high level of formal accuracy, as the figures of 90% and 94% in the last row of Table 5 indicate. While AWL words used in such sentences as "Life today is manipulating us and we can deny that because most of us," exhibit morphological correctness, such sentences are clearly not very meaningful. It is more interesting to consider the possible sources of Hussein's gains in semantic accuracy, which increased a great deal -- from 40% to 61%.

In his response to the open-ended question about other ways of studying, Hussein noted his strategy of "doing the words on the AWL quiz." This refers to using the quizzing function of the Word Bank rather than gapped-concordance self-quizzes. These quizzes require the user to supply a missing word in a single sentence example that has been entered by a fellow student rather than drawn from a corpus of texts designed for native speakers. This practice strategy may have proved to be a more manageable task for this low proficiency learner and may help account for his impressive gains in definitional knowledge. It appears that Hussein may have been more unable than unwilling to use concordances based on authentic academic texts.

	Pre-course	Post-course
Vocabulary Levels Test (%)	73.00	87.00
Text: No. of AWL items	10	18
Text: Meaning accuracy (%)	40.00	61.11
Text: Formal accuracy (%)	90.00	94.44

Table 5: Profile of scores and AWL use for Hussein (non-user)

Discussion

In this study, we have seen that many students in an experimental vocabulary course made regular use of a concordancer and other tools in their study of AWL words, though there was also considerable variability. Purposes for concordancing varied; there appeared to be a slight tendency for students to use the tool to "look at examples" more often than they used it to find collocations or make gapped quizzes. The students were enthusiastic about the materials used to teach vocabulary in the course and the new words learned. Group means for performance on the AWL section of the Vocabulary Levels Test indicate that they made substantial gains in their ability to recognize the meanings of words targeted in the course. Pre/post comparisons of learner essays also indicated growth in ability to use AWL words productively.

To what extent can this growth be attributed to concordancing activities? At the beginning of this chapter, I noted the difficulty of substantiating the link between a study strategy such as concordancing and the linguistic gains learners may achieve. In the present study, it seems likely that the combination of a dedicated teacher and a textbook that the students clearly valued accounted for a great deal of the learning they achieved. It is likely that students' use of traditional, non-computerized review techniques also contributed to their retention of AWL words. Nonetheless, the profiles of users and non-users indicate that concordancing may play a critical role in the development of ability to use new words productively. It is striking that the two avid concordancers, Jamal and Sadaf, were both able to incorporate many more words in their writing (a demanding task) at the end of the course than at the beginning. In both cases, figures are almost double. Moreover, both learners were able to use the new words with high levels of semantic and formal accuracy.

The profile of the non-user, Yuan, also speaks to the potential usefulness of concordancing. By his own testimony, he is prepared to "cram" and "study hard". But Yuan's writing performance indicates that knowing more word meanings is not really what he needs. More attention to the formal characteristics of words in use – an aspect that the concordancer highlights as perhaps no other medium can – would likely benefit this learner a great deal. The case of non-user Hussein speaks to the challenge of motivating students to use concordancing. This student appeared to be willing to use other computer tools (e.g. the Word Bank quiz feature) but was not inclined to use the on-line concordancing options that were available. The question of motivation is one of

several issues that need further investigation; directions for future research are outlined in the next section.

Conclusion

Motivating students to use on-line concordancing resources remains a challenge for designers of computer-assisted vocabulary instruction. Although previous investigations of concordancing (Cobb, 1997; Cobb,1999) point to word knowledge gains that are qualitatively different from those achieved when students use more traditional study techniques, the rewards of study using concordancing activities may seem intangible to many learners. The learners in this study clearly preferred to use computers in other ways. As reported above, the most popular of the on-line tools proved to be an on-line dictionary -- a resource that offers immediate rewards in the form of answers to quests for information. This preference for on-line dictionaries was also found in earlier investigations of on-line tools for vocabulary learning (Cobb and Horst, 2001; Horst, Cobb & Nicolae, 2005). The finding raises the question of whether concordancing might be more popular with students if it were 'sold' in the same way that dictionary use is. That is, rather than promoting concordancing as a way of studying for vocabulary quizzes, ESL teachers could encourage students to consult a concordance for information about words in use, much as they would consult a dictionary for information about meanings. Concordance searches based on genuine need to use new words accurately in writing tasks might well create conditions for highly effective data-driven learning. This is an avenue for further research; a step in this direction has been taken by Plomer (2004), who is investigating the effects of access to an on-line concordancer in a task that requires ESL learners to produce written narratives.

Ideally, future experimentation will also involve larger numbers of participants so that links between word learning gains and use of concordancing (and other strategies) can be tested statistically. The small size of the group available for this experiment necessitated a case study methodology. While this approach succeeds in providing insights into students' use of the computer tools and the vocabulary growth they experienced, it does not provide a firm basis for claiming that concordancing played a determining role. Future experiments would also benefit from more accurate measures of learners' use of the on-line tools. Rather than relying on self-report responses as in the research reported here, future studies might track amounts of time participants actually spend using the various options. In this study, an attempt to track participants' out-of-class tool use was made by collecting IP numbers of most often used computers and adding to the class website a log-on box that required participants to identify themselves by name. However, sophisticated computer users in the group bypassed the log-on and accessed the tools directly (all available at www.lextutor.ca) with the result that user protocols did not reflect use levels accurately. Future research can avoid this problem by conducting the experimentation in a computer lab where students' use of the resources can be closely observed.

Finally, I return to the unmotivated concordancer, Hussein; his case highlights an issue that is crucial to the success of concordance-assisted vocabulary instruction in academic ESL settings. Hussein's lack of interest in the concordancing option likely is because he

knows so few of the vocabulary in the concordance lines. Given that his knowledge of AWL vocabulary is low, it is reasonable to assume that his knowledge of low frequency English words is also limited, making the language of the concordance lines very difficult for him to comprehend. Indeed, many native speakers might have difficulty inferring the meaning of *construction* if success depended on their understanding of terms like *butterfly diagrams, honeycomb panels* and *metalogs* (see concordance of *construction*, Figure 1, lines 14, 15 and 16). Clearly, there is a need for an authentic but learner-friendly academic corpus that can support study of the AWL. The texts that make up this corpus should be selected with a view to avoid confronting ESL learners with undue occurrences of the *honeycomb panels* and *metalogs, which are elicited from* native speaker-oriented corpora now available. The potential of concordancing to benefit second language vocabulary acquisition is by now well established. The development of a corpus that ensures learners can reliably access comprehensible contextual information will be a useful step toward making the benefits of concordancing more widely available for all.

Acknowledgements

This research was made possible by funding from Fonds québécois de la recherche sur la société et la culture (FQRSC). I would like to thank research assistants Roisin Dewart, Carol Johnson, Angela Kleanthous and Jeff Millar -- all students at the TESL Centre in Concordia University's Department of Education for their help with this project.

Endnotes

1. Spelling errors in AWL words proved to be rare. Where they occurred, the writer's intention was clear (e.g. *enviromental).
2. Cases in which one of the two raters judged a word use to be acceptable were treated positively in this study. That is, a single rating was seen as an indicator of acceptable (though perhaps not universally acceptable) usage. When performance was analyzed using a stricter both-only criterion that awarded no points to words judged acceptable by one rater only, patterns in the results proved to be identical to those reported here (although differences were less pronounced).

3. Readers may wonder why formal accuracy percentages are higher than meaning accuracy percentages, as can be seen in the case of Jamal's pre-course writing sample in which 75% of semantic uses and 83.33% of the formal uses are correct. This occurs because learners are able to use words that they do not know semantically in ways that are formally correct. For instance, Jamal writes, "sources of the earth like water, oceans, etc.," The use of the word *sources* here is semantically incorrect but it is morphologically and collocationally accurate: the learner clearly knows that *source* is a count noun that can be pluralized using *s* and be followed by *of*. The rating scheme was designed to be sensitive to partially acquired vocabulary knowledge in this way. In this case, the learner appears to have mastered formal aspects of the word's use in sentences but not yet made the semantic distinction between *source* and *resource*.

4. Yuan's results may be coloured by lack of motivation to complete tasks on the post-course measures. This (rather than lack of linguistic ability) may explain his use of just nine AWL words in the post-course sample in contrast to the 14 used earlier and his slightly lower post-course score on the Vocabulary Levels test.

5. The class teacher reports that Hussein has a high level of oral proficiency. For this reason, the characterization of Hussein as low-proficiency learner may be somewhat inaccurate. He appears to have acquired the oral skills needed for everyday interactions but has not yet mastered academic discourse, which contains much larger proportions of AWL vocabulary than normal speech (Nation, 2001). Hussein's ability to speak accurately may explain his high scores for formal use of words whose meanings he clearly does not understand.

References

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System,* 25, 301-315.

Cobb, T. (1999). Breadth and depth of vocabulary acquisition with hands-on concordancing. *Computer Assisted Language Learning, 12*, 345-360.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly,* 34, 213-238.

Cobb, T. & Horst, M. (2001). Growing academic vocabulary with a collaborative online database. In B. Morrison, D. Gardner, K. Keobke, & M. Spratt (Eds.), *ELT perspectives on IT & multimedia; Selected papers from the ITMELT Conference 2001* (pp. 189-225). Hong Kong: Hong Kong Polytechnic University.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with a collaborative on-line database. *Language Learning and Technology, 9,* 90-110.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement, *Applied Linguistics,* 22, 1-26.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Plomer, M. (2004). *The use of the concordancer as a complement to the on-line dictionary in EAP writing*. Unpublished MA thesis proposal. Concordia University, Montreal.

Schmitt, D., & Schmitt, N. (2005). *Focus on vocabulary: Mastering the Academic Word List*. New York: Longman/Pearson Education.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

Contributors

Tom Cobb is a professor of TESL and applied linguistics at the University of Quebec at Montreal (UQAM). He was an ESL instructor, coordinator, and course designer for several years in Europe, the Middle East, and Hong Kong; he did his PhD in educational technology at Concordia University in Montreal.

Robert Courchêne is an ESL teacher and teacher trainer attached to the Second Language Institute of the University of Ottawa. His research interests include culture and multiculturalism as they relate to teacher training, testing and curriculum design. He haired the project that produced the *Standards lingistiques canadiens* (Benchmarks in French) and the accompanying placement test.

Janna Fox is an Assistant Professor within the School of Linguistics and Applied Language Studies, Carleton University, where she teaches courses in curriculum, language testing and research methodologies. Her research interests include language testing, curricular innovation, and the interplay of curriculum testing and classroom practice.

Marlise Horst is an assistant professor in the TESL Centre at Concordia University. Her research focuses on second language vocabulary acquisition. She teaches courses in methodology, language testing and the history of English. Before joining Concordia, she taught ESL in a variety of international settings including Egypt, Oman and Hong Kong.

Gail Stewart is an educational consultant whose work in assessment includes the CLBA, CLBLA, CLBPT, MLPT, COPE, and CAMERA. She is a former ESL instructor and teacher trainer in the University of Toronto's Intensive ESL and TESL Certificate Programs, Faculty of Education, and School of Graduate Studies.