Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies

Tom Cobb

Abstract: The characterization of learner interlanguage has been largely confined to the early acquisition of speech in a second language (L2), with later acquisition not commonly analyzed in this framework for lack of theory, lack of data, and the fact that late acquisition is intermeshed with the acquisition of literacy. The current trend to electronic submission of classroom writing relieves the data problem, defines steps in the acquisition of literacy, and may even contribute to the growth of theory. When assembled by teachers or researchers into a learner corpus (LC) of suitable size and character, such a corpus provides the empirical means to discover what advanced learners know and do not know about their L2. A strong tradition of LC analysis has emerged in Europe; the present study introduces this work and tests its applicability to a North American context.

Résumé : La caractérisation de l'interlangue des apprenants a surtout porté sur l'acquisition précoce d'une langue seconde à l'oral en raison de l'absence de théories et de données permettant d'analyser le phénomène de l'acquisition tardive, cette dernière se confondant d'ailleurs avec l'acquisition de la langue écrite. La tendance actuelle consistant à soumettre par voie électronique les textes rédigés dans le cadre des cours pourrait offrir une façon de pallier le manque de données et de définir les étapes de l'acquisition de l'écrit et pourrait même mener à l'élaboration de théories. Les enseignants et les chercheurs qui ont à leur disposition un corpus adéquat quant au nombre de textes et au caractère de ces derniers possèdent les moyens empiriques de découvrir ce que les apprenants avancés savent et ne savent pas de leur langue seconde. Il existe une forte tradition d'analyse de corpus de textes d'apprenants en Europe. Cette étude des travaux qui en sont représentatifs examine les conditions de leur application dans le contexte nord-américain.

Introduction

A concept that has held up well from 1970s second language acquisition (SLA) theory development is interlanguage (IL; Selinker, 1972), according to which learner language displays systematicity and opportunity for intelligent intervention rather than random error. An IL framework has been useful in understanding early second language (L2) acquisition (e.g., Dulay & Burt, 1974) but has contributed less to understanding the continuing development of intermediate and advanced learners. Too often, IL analysis of later acquisition revolves around the concept of fossilization, a notion only slightly more illuminating than *error*. The continuing need for 'IL-sensitive research methods' has been noted by, among others, Larsen-Freeman (personal communication, 2000).

There are two reasons that intermediate-advanced IL remains relatively uncharted, the first being lack of data. When school children are beginning to communicate in L2, it is fairly simple to tape-record and transcribe their classroom interactions over time and then compare these either to a stage theory of acquisition (Brown, 1973; Pienemann, 1999) or to features of the children's classroom instruction (Bloom, Hood, & Lightbown, 1974). Children typically start from a common baseline and are available for further study on a continuing basis. Intermediate and advanced learners, on the other hand, tend to be older, more diverse in start point and study mode, and less available for extended observation. Also, much of the development of advanced learners involves learning to read and write in the L2, and interaction with written texts is less amenable to data capture than spoken interaction among learners (text comprehension is all but impossible to observe naturalistically).

Text production, however, is a potentially rich source of information about advanced IL development. The problem with using written production as IL data has been, until recently, the practical one of finding a way to collect and analyze learner texts in any volume. However, with increasing numbers of on-line submissions in schools and universities, huge stores of potential IL data are becoming available.

But even if we collected thousands of machine readable texts of advanced learners we might not know what we were looking for – other than native performance or its absence, i.e. error. This is the second reason that advanced IL remains uncharted, that we have no strong theories of late acquisition comparable to, say, the theories of morpheme acquisition. Even if commonalities in advanced learning exist, these are likely to be in areas beyond morpho-syntax, such as lexis, discourse, and

pragmatics, where we have few characterizations of native speaker (NS) performance, let alone of the sequence(s) by which learners arrive there. So while the shortage of data is no longer a problem, the shortage of theory remains. It is, of course, conceivable that these two problems will be resolved in tandem, in other words that along with the accumulation and study of advanced learner data, hypotheses and theories will emerge. This paper looks at three attempts to develop and test hypotheses about advanced IL using computerized learner text as its evidence.

Several researchers in both linguistics and applied linguistics are now devoting significant career time to devising methods of analyzing and interpreting large bodies of electronic writing (text corpora) produced by both NSs and learners, with a view to characterizing NS competence and tracing sequences toward it. Corpus linguists use large corpora to provide a more detailed and accurate description of the lexis, discourse, pragmatics, and of course morpho-syntax of NS English than has been previously available (Biber, Conrad, & Reppen, 1998; Sinclair, 1991; Stubbs, 1996). Similarly, applied corpus linguists use corpora of learner writing to put together a description of the stages learners go through as they move toward NS competence in these same areas. Particularly promising is the comparison of the two corpus types, which can reveal not only what *is* in a learner corpus, but also what is *not* in it. A notable moment in the development of this new methodology of contrastive corpus analysis is Sylviane Granger's (1998) *Learner English on Computer*, which describes several pioneering efforts in Europe to collect and interpret learner corpora.

An interesting approach adopted by several of Granger's contributors is to work from a common observation or impression about advanced learner language, develop a hypothesis to explain the observation, and test the hypothesis through a comparison of learner and NS corpora. For example, it is often observed that learner writing even at apparently native level remains nonetheless 'vague,' or resembles NS speech written down more than it does NS writing. Three hypotheses have been proposed to explain this impression. One is that such learners rely on the restricted, context-determined lexicon of spoken language rather than deploying the broader lexicon typical of NS writing. Another is that the lexicons of NS and advanced learners are similar, but the phrase structures are not. A third explanation is that NS texts of equivalent genre display less personal involvement than learner texts do.

The present study will provide an introduction to the European work by showing how these three hypotheses have been investigated through contrastive corpus analysis. Further, it will test each of the European

findings against a recently collected corpus of Québec learner writing (see materials section below), with the aim of discovering whether there is a common pattern of interlanguage development across relatively distinct populations of advanced learners. As a reviewer of this paper has noted, 'Corpus linguists regularly point out that one of the advantages of corpus-based studies is that they are replicable ... however, very few replication studies have been carried out so far.' Since each investigation and replication will be dealt with in turn, the literature review will unfold over the course of the report. Similarly, each of the studies raises characteristic method and technology concerns which will be treated as they occur. Many of the tools and materials mentioned in the study are available on the author's *Compleat Lexical Tutor* web site [a]. (Web site addresses are indicated by square brackets and appear as a separate set of references at the end.)

There are practical benefits to having a better description of advanced IL. Until recently the main advice that language teachers could give advanced learners was 'to get lots of practice,' assuming there was little more that focused instruction could do for them. These learners have become, as it were, defective native speakers, working with restricted linguistic resources and dealing with a remainder of highly individual random errors. If, instead, advanced learners are seen as learners nonetheless, moving systematically through acquisition sequences and overcoming shared misconceptions about the L2, then instruction can be focused more effectively throughout the learning process. Each summary and replication below will end with a speculation on pedagogical implications.[1]

Materials

Text corpora

It is a common misconception that corpus building means collecting lots of texts from the Internet and pasting them all together. In fact, corpus building is a large and complex topic in its own right (see McEnery & Wilson, 1996, on 'corpora vs. machine readable texts'). Decisions must be made about the type of materials to collect: Will they be language general or domain specific, spoken or written, native or learner? Each type of corpus has its own inclusion rules and sampling procedures (see on-line documentation [b] describing the composition of the early million-word Brown corpus [1972], and the *Lexical Tutor* web site [a1, a5] for examples of its output). Additional rules apply to corpora intended for comparison, namely that the corpora be of similar size and produced

under similar circumstances. Most of the learner corpora (LCs) used in Granger (1998) are between 50,000 and 150,000 words and are the result of typical academic, expository writing assignments produced by learners with a common L1. Ideally, the NS corpora used in these studies are the result of similar writing tasks performed by equivalent groups of NSs, but often these are not available and general corpora of NS writing are used instead.

The Québec learner corpus to date consists of over 250,000 words, divided in two main sections, advanced and intermediate ESL learner writing. The main section of interest is the advanced learner corpus, which is advanced in the sense that it was produced by non-native speakers of English who had been successful applicants to a TESL training program at the Université du Québec à Montréal (UQAM). These students (n>400) had taken a computer based admission test between 1997 and 1999 with a writing task of 150–200 words on the expository topic 'How could English teaching in Québec be improved?' and the corpus (henceforth the TESL corpus) is a collection of the essays of the students admitted to the program (about 80% of applicants).

The intermediate LC comprises roughly 100-word placement test essays on the topic 'What difference would it make to your work or life if your English was significantly better than it is now?' written by more than 1500 students applying for ESL courses at the same institution. This corpus (henceforth the ESL corpus) is divided by overall score on the placement test into High, Medium, and Beginner levels. The beginner corpus is smaller than the others since there are officially no beginners' courses in English offered at UQAM. In the present study, these ESL corpora serve mainly as supporting documents in that they will be explored for the emergence of trends in the TESL corpus.[2]

All learner corpora are untreated except for spelling correction. Baseline NS corpora are the same ones used in the original European studies where possible, with any additional NS data taken as needed from the Brown corpus, the British National (BNC) written or spoken corpus (one million words apiece) or learner-appropriate textbooks and other materials, all of which can all be sampled on the Lexical Tutor web site. Corpus parts and sizes are summarized in Table 1.

Software for corpus analysis

When a corpus is large enough to provide interesting information, it is too large to be interpreted without the aid of computational tools. Tools have been developed which hold enormous quantities of information in memory and produce such information as word counts, frequencies,

TABLE 1
Corpora and sizes

TESL Corpus	
English teacher trainees	75,437 words
ESL Corpus	
High	70,015 words
Medium	70,375 words
Beginners	37,439 words
Total	253,266 words

and collocational and syntactic patterns. The most widely used tool for general corpus analysis is Mike Smith's Wordsmith [c], and other tools have been developed for specific tasks (such as Paul Nation's VocabProfile [d], which breaks texts into components by lexical frequency). Rudimentary versions of these and other tools of textual analysis can be found on-line at the *Lexical Tutor* web site [a1, a2, a3].

European Studies and Québec replications

In the main body of this paper, three European corpus comparison studies are reviewed, their findings are replicated with a corpus of Québec learner writing, and pedagogical implications are proposed.

Study 1

H. Ringbom, *Vocabulary Frequencies in Advanced Learner English: A Cross-linguistic Approach* (pp. 41–52).

Summary

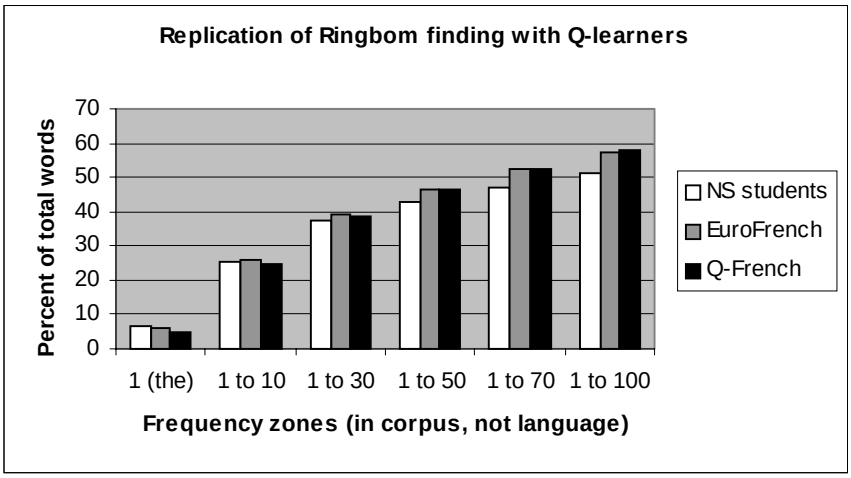
As already mentioned, a common approach in learner corpus (LC) studies is to begin with an intuitive observation of advanced learner language, propose a hypothesis to explain the observation, and then test the hypothesis with an NS/NNS corpus comparison. The following quotation from Ringbom's chapter encapsulates the approach:

> A frequently voiced view is that learner language is vague and stereotyped. This would be a natural consequence of its vocabulary being more limited than that of native speakers. However, concrete evidence of exactly what constitutes this vagueness has been hard to come by. (p. 49)

The observation is that advanced learner language is vague and stereotyped, and the proposed explanation is that vagueness stems from lack of vocabulary. The empirical evidence Ringbom uses to support this hypothesis is a comparison of the use of the 100 most frequent words of English across a set of learner corpora for seven L1s and comparable native speakers. The 100 most frequent words are of course mainly pronouns and other function words (see [e] for a comparable list or [f] for a complete range of frequency lists). Over-use of basic vocabulary indicates, of course, under-use of other, richer, more precise, and more varied vocabulary. Ringbom examines learner reliance on the 10 most frequent words, 20 most frequent, and so on, up to the 100 most frequent.

His finding is that advanced learners across seven L1 backgrounds consistently use these 100 very high frequency words in their writing about 4–5% more than NS writers (see Figure 1 below). It is not the very highest frequency function words that they overuse ('the,' 'in,' 'of,' or other top-ten items), because these tend to appear mainly in obligatory contexts. Rather, it is the slightly less common function words from the 30–100 zone ('which,' 'into,' 'because,' 'about') that are overused, along with some very common content words ('people,' 'new,' 'many,' 'different,' 'important') where numerous variants are possible. For example, advanced learners use '(I) think' between three and five times as much as NSs do, presumably for lack of confidence with alternatives

FIGURE 1
Ringbom's overuse finding replicated



like ‘judge,’ ‘believe,’ and ‘consider,’ equivalents of which exist in all the learners’ L1s.

It is plausible that repetition of high frequency items and failure to nuance common notions may well account for the sense of vagueness that native speakers find in advanced learner writing. Admittedly, the evidence is merely correlational: there is vagueness, and there is overuse of high frequency lexis, but no causal connection is actually established. In fact, a correlational underpinning typifies much of the current LC research and may be related to the novelty of the approach. The next step in the research agenda is presumably experimental hypothesis testing. Here, for example, Ringbom might have gone on to empirically test teachers' vagueness ratings against learner texts of varying lexical density, or regressed vagueness ratings against several candidate factors. LC research amounts to a new paradigm, and a great deal of methodological pioneering remains to be done. In the replications to follow, however, no attempt is made to press beyond the correlational phase.

Replication and extension of Ringbom's finding

Replication

Few of the studies in Granger (1998) are sufficiently detailed to make it totally clear how one would go about a replication, once in possession of a corpus, but the following is the method that was used here. Ringbom had looked at the proportion of advanced learner writing that was accounted for by words at each tenth percentile up to the 100 most frequent words of English. To get a comparable result, the Wordsmith program was used to break the advanced Québec LC into a frequency list – a list of all the words in a text in order of frequency, also indicating the percentage of the corpus that each item accounted for (similar software is now available on the Lexical Tutor, although limited to the Internet text input of about 30,000 characters [a3]). A manual sum of percentages was taken after every ten words (see Table 2 for an example of how this was done for the first ten words, which account for just under 25% of all lexical items in the Québec LC).

The question of interest is whether 25% for 10 words is a little or a lot, and so on down the frequency list. As Figure 1 makes clear, the European Francophone learners’ use of the first 10 words is about the same as for comparable NS writers (mainly obligatory contexts), but for items 30 to 100 (mainly very general content words like ‘people’ and ‘think’) there seems to be a pattern of overuse, a pattern reflected almost identically in the North American Francophone LC.[3]

TABLE 2
Method of calculating use of very high frequency vocabulary in Quebec advanced LC

N	Word	Frequency in corpus	% of Corpus
1	the	3,503	4.65
2	to	2,891	3.57
3	I	2,177	2.89
4	a	1,780	2.38
5	of	1,702	2.28
6	and	1,632	2.17
7	in	1,485	1.97
8	that	1,478	1.96
9	is	1,265	1.68
10	it	990	1.31
			24.82%

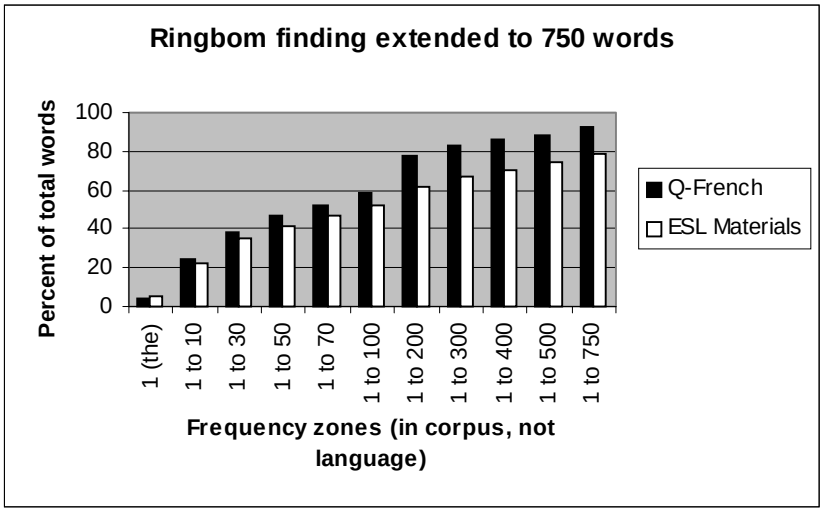
Extension

Two interesting questions follow from Ringbom's study, which are whether the pattern of overuse continues to common words beyond the 100 most common, and whether the pattern of overuse changes with time and increased proficiency. For the second question, the value of a collecting a graded set of learner corpora will become apparent.

To investigate the first question simply required extending the method already described further down the frequency list but at larger intervals. An arbitrary cut-off was chosen of the most frequent 750 words. A new NS corpus was needed for this extension of the study, since there was no comparison data for these frequency levels in the original study. For this, a 63,000-plus word corpus of ESL instructional materials written by native speakers and originally developed as a corpus for another purpose (Cobb, 1997) was used. This corpus of ESL materials was used as a way of giving maximum advantage to the LC, since several studies have shown such materials to be lexically basic (Meara, 1993; Cobb, 1995), with a preponderance of their lexis drawn from the first 1000 words (about 90% as opposed to the usual 70%, see below). The answer to the first question about common words is that the tendency to overuse high frequency items not only continues but increases. In Figure 2, it seems clear that a roughly 5% overuse factor up to the 100 word mark gives way to a roughly 10% factor thereafter. (It should be noted, however, that the jump corresponds to a change of comparison corpus, a potentially confounding factor.)

The items from 100 to 750 are exclusively content words, again mainly of a general nature. Samples from the frequency list for the BNC

FIGURE 2
Overuse of common items, extended findings



[f] at roughly 100 word intervals include word 200 'things,' 302 'problem,' 402 'position,' 502 'change,' 600 'strong,' and 702 'everyone.' It is interesting that these very general terms are overused while far more nuanced versions of the same basic concepts inhabit only slightly remoter frequency zones. For instance, 'difficulty' for 'problem' is at frequency position 771; 'powerful,' 'solid,' and 'dynamic' for 'strong' are at 1341, 2455, and 4085, respectively. So once again we find evidence of advanced learners overusing general, unnuanced lexical items.

Another way of testing and possibly broadening the overuse hypothesis is to switch to a slightly different question and a different computer tool. The tool is Nation's VocabProfile (Laufer & Nation, 1995), which deconstructs any text or corpus into its lexical components by frequency, indicating the number of words from the 1–1000 frequency zone, from the 1001–2000 zone, from the Academic Word List (AWL; Coxhead, 2001; [g]), and finally from beyond all three previous zones. Different genres of text often have distinct profiles according to this scheme, as can be seen by analyzing profiling the sample texts provided with the on-line version of the program [a2] with accompanying tools for Chi-square comparison. Typical Vocabprofile output for written and spoken NS English are shown in Table 3. Of main interest is the 0–1000 zone of high frequency items, which typically comprise 70% of written NS expository texts and 80% of spoken conversational texts. The

TABLE 3
Lexical profile of NS writing, NS speech, and advanced learner writing

	NS student writing	NS speech	Advanced learner writing
0–1000	70%	80%	88 %
1001–2000	10%	5%	3 %
AWL	10%	5%	3 %
Off-list	10%	10%	6 %

common explanation of the written-spoken difference is that spoken language, especially conversation, does not require nuanced vocabulary since nuancing of meaning can be provided by shared context, deixis, facial expression, and so on. Most forms of writing, on the other hand, have greater need of nuanced vocabulary since written texts must be able to bridge gaps over space and time between unshared contexts. (Admittedly, certain forms of poetry specialize in wringing fresh meanings from worn words so that not every instance of language production can be illuminated by frequency analysis—*To be or not to be* is the classic example of this.) If the Québec LC is fed through Vocabprofile, will a general pattern of overuse be found in the 0–1000 zone as a whole?

The result of this analysis is that almost 90% of vocabulary items used in writing by these advanced learners are common words from the 0–1000 frequency range. In other words, when required to perform a high-stakes writing task on an objective topic for an anonymous reader, these learners simply employ the restricted lexicon of speech, 'writing down talk' as it were. Once again, an overuse hypothesis is confirmed.

The second extension of Ringbom's study involves investigating whether the overuse of basic vocabulary decreases over time, and if so how much and how fast. To answer this question, one would ideally have recourse to large writing samples from these same or equivalent learners over the years of their remaining studies, and indeed such an evolution in corpus building is currently underway (with the difficulties of tracking advanced learners as already noted). Meantime, the ESL corpus of learner writing at three levels can be used to experiment with methods and provide some indication of what might be found. Extrapolation from cross-sectional to longitudinal data is a characteristic of LC methodology, as it was in earlier interlanguage studies.

VocabProfile analysis shows the ESL and TESL corpora to be about equivalent at the 0–1000 level, both employing about 90% of items from this zone, although this similarity probably represents a ceiling effect as well as masking some interesting differences. To locate these differences in a principled manner, we return to the notion that high frequency

vocabulary is mainly unnuanced vocabulary – the overuse of 'think' (frequency rank 64) when items like 'believe' (273), 'consider' (349) or 'suspect' (2218) would be more appropriate, and so on. (See [a2] or [a5] for the means to test these or other word frequency ratings.) Does 'think' gradually differentiate into *believe* and other nuanced alternatives as students move through the levels of language learning? A way of finding out is to check for the occurrence of several basic concepts and their less frequent nuanced versions through corpora of comparable learners at several levels.

A problem involved in searching corpora for single items is of course that such items are only of interest if spread to some extent throughout the corpus. For this Wordsmith's dispersion features can be used to check whether the use of a particular word or phrase is spread throughout a corpus or piled up in one corner of it. Figure 3 shows such a dispersion plot for instances of 'believe' across the three ESL corpora, the TESL teacher corpus, and two NS corpora (the ESL materials already mentioned, and a newspaper corpus supplied by John Milton [h]). It seems clear that 'believe' is more frequent as one moves up the levels as well as more generally dispersed (even in the advanced or teach corpus there are dumps of black marks with open spaces between, compared to the steadier rhythm of the two NS corpora).

'Think' and 'believe' and several other less-more nuanced pairings were extracted from the four learner sub-corpora, subjected to the dispersion check, and plotted on graphs so that any patterns would become visible. Examining the frequencies of these items across the sub-corpora reveals quite clearly that several common, all-purpose items gradually recede and make way for more nuanced alternatives. In all items tested, the finding is similar to 'think' and 'believe' in Figure 4 (including 'because' and 'as a result of,' 'in' and 'into,' and several others which can be further investigated by the reader in on-line materials [i]). The common item gradually decreases, the nuanced alternative gradually emerges. Interestingly, this crossover phenomenon probably shows these learners crossing a threshold (in the sense of Alderson, 1984) where they will have access to a range of resources similar to those they deploy in L1--literally, in the case of 'believe' and 'think,' since 'je crois' ('I believe') is the normal form of 'I think' in the L1 of these learners. To summarize, the overuse seems to decline with time and greater proficiency, although slowly.

Extending Ringbom's investigation, then, we have seen that the overuse phenomenon is probably even more persistent, in Québec at least, than the original analysis had shown, extending well beyond just the first 100 words and corresponding in more general terms to the

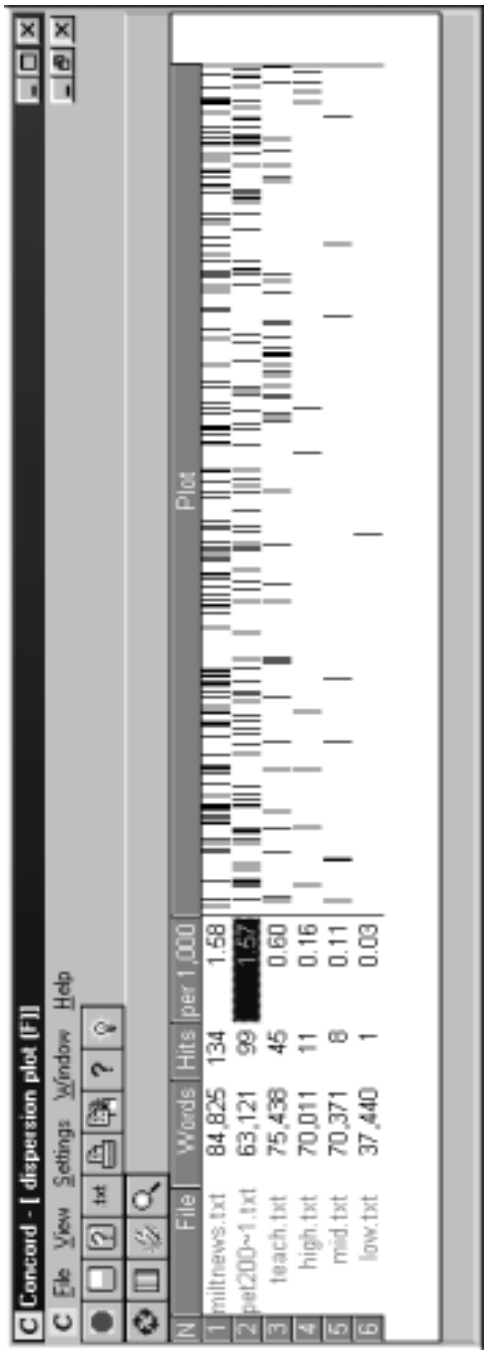
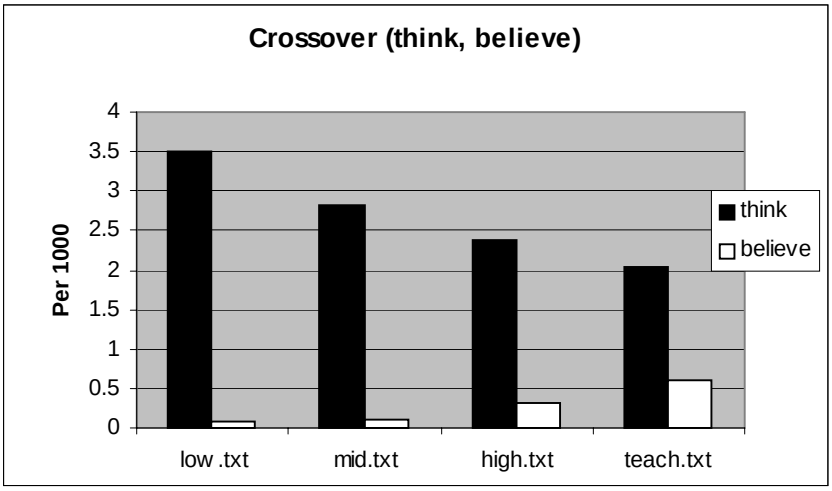


FIGURE 3
Wordsmith's dispersion plot for *believe*

FIGURE 4
Gradual differentiation



restricted lexis of spoken language. On the bright side, we have also seen learners steadily working free of this lexical confinement. The working free is, however, extremely gradual, amounting to fewer than .6 instances of 'believe' per 1000 words in the case of TESL trainees vs. more than 1.5 in two independent NS corpora (Figure 3). A graded LC, in other words, discloses the slow emergence of a native-like repertoire of vocabulary resources that could almost certainly be hastened by effective instruction.

Pedagogy

Vocabulary instruction in advanced courses, where it exists, typically consists of piling more and more low-frequency vocabulary into learners' heads for passive use in reading comprehension. The findings presented here suggest that some attention should also be paid to two other areas:

1 Some effort might usefully be devoted to the diversification of very high frequency items ('think,' 'in') into the more nuanced and only slightly less frequent items ('believe,' 'into') which allow for both greater differentiation and more native-like production.
2 Vocabulary courses should include the teaching of vocabulary for productive use since, at least in the Québec setting, these learners'

scores on a test of recognition vocabulary (Nation's Vocabulary Levels Test, 1990) typically show their vocabulary knowledge to be far greater than what comes through in their writing. In a study involving vocabulary testing of these same learners, Cobb (2001) found substantial passive vocabulary knowledge at levels well beyond the most basic.

Study 2

S. de Cock, S. Granger, G. Leech, & T. McEnery, *An Automated Approach to the Phrasicon of EFL Learners* (pp. 67–79).

Summary

Once again the European study starts from a familiar observation about advanced learner language, in this case that even when this language is largely free of errors it remains nonetheless 'foreign sounding.' One hypothesis that has been advanced to explain this (Kjellmer, 1991, p. 124) is that in these learners' production, 'the building material is individual bricks [words] rather than prefabricated sections [lexicalised phrases].' In other words, advanced learners operate on the open choice principle of language production (where, for example, any noun in the user's lexicon is equally eligible when NP is the forthcoming syntagm), while NS's operate to a greater extent on what Sinclair (1991) called the idiom principle (where some nouns are more likely than others to fit in particular environments).

The background to this distinction is the growing consensus that fluent language production, particularly oral production unfolding in real time, would be impossible if each syntactic choice point had to be negotiated creatively (in the Chomskyan sense), and that NSs instead rely heavily on multiword items (MWIs) or prefabs (precast phrases like 'by the way' and 'if you see what I mean') that allow for coasting and save cognitive resources for important choice points. Prefabs allow several words to be purchased for the price of one, in terms of cognitive economy, creating savings which are invested in discourse planning, memory search, and the like (Pawley & Syder, 1983; Moon, 1997; Wray, 2001). Kjellmer's guess about the foreign-soundingness of advanced learner English is that it stems from a much reduced incidence of prefabs.

De Cock et al., test Kjellmer's idea against a matched set of 25 transcribed learner (French) and NS university admission interviews. The two corpora were run through McEnery's automatic phase extraction

program Tuples, which extracts from a text all recurrent word combinations of a given length and frequency. Tuples' output is thus two long lists of all the recurrent phrases in the two corpora ('by the way' but also 'in the big,' and so on) accompanied by counts and basic statistics.

Despite the plausibility of Kjellmer's hypothesis, the finding of de Cock et al., goes against it to some extent. It turns out that advanced learners do use precast phrases, in fact use them more than native speakers do. On reflection this is no surprise, if it is true that fluent language production would not be possible without relying on precast phrases – most advanced learners being reasonably fluent. However, what distinguishes learners from NSs, these researchers find, is the small number of precasts advanced learners have at their disposal, and the extent to which these are used and overused. This finding, then, is similar to Ringbom's finding in the previous section: advanced learners' phrases, like their words, are few in number and overused.

Replication

Since the program Tuples was publicly unavailable at time of writing, Smith's program Wordsmith [c], which has an automatic phrase extraction or cluster tool, was used to gather and tally all the MWI's or recurring strings from both sections of the Québec LC. The European and Québec corpora are comparable in that both subject matters are constrained (admission interviews in the one and a view on English teaching in Québec in the other), and hence could be predicted to contain some amount of lexical repetition, although not necessarily phrase repetition. However, the two corpora are not comparable in that the Québec corpus is a written corpus while the European corpus is transcribed speech (the phrase research has been conducted mainly in the context of speech production). With this proviso, the preliminary finding is essentially the same on both sides of the Atlantic. Table 4a shows raw Wordsmith output for the mid-level ESL learner sub-corpus, with the five most frequently occurring two and three word strings. As the table shows, the most frequent two word string is 'my English.' Percentages are given for each string as a proportion of all strings of the same length. Table 4b shows assembled output for all two-word and then three-word strings across a set of learner and native corpora (once again, News and ESL materials).

Table 4b summarizes the degree of repetition of the 100 most frequent two and three word strings across two native and three learner corpora. All figures are percentages, expressing how largely each string features across equal length strings throughout the corpus. For example, the first

TABLE 4a
Most repeated clusters

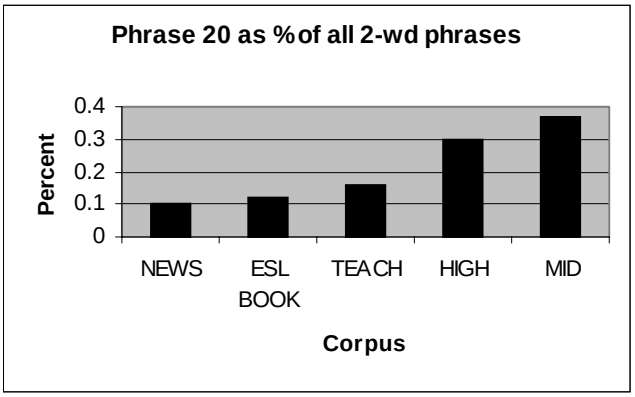
N	Two word	Freq	%	N	Three word	Freq	%
1	My English	447	.64	1	I want to	293	.42
2	English is	403	.58	2	I have to	199	.28
3	Want to	369	.53	3	If my English	197	.28
4	I want	358	.51	4	Be able to	184	.26
5	I have	356	.51	5	My English was	151	.22

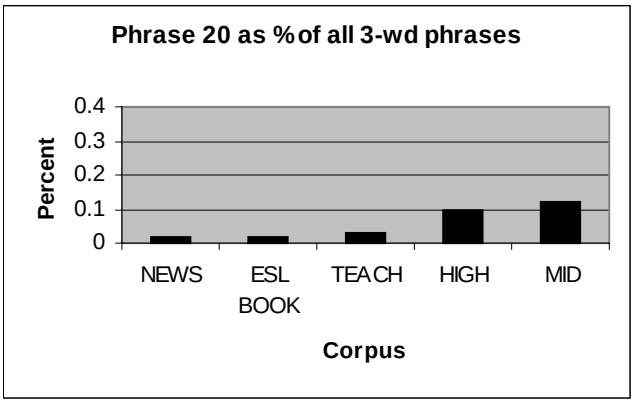
TABLE 4b
Phrase repetition as percentage of total same-length phrases

				ESL						ESL	
	News	Materials	TESL	High	Mid		News	Materials	TESL	High	Mid
2-WD						**3-WD**					
String						String					
1	.58%	.57%	.41%	.70%	.64%	1	.05%	.15%	.13%	.33%	.42%
2	.28	.44	.38	.64	.58	2	.04	.06	.10	.29	.28
3	.21	.35	.32	.54	.53	3	.04	.04	.09	.25	.28
4	.18	.22	.31	.47	.51	4	.02	.04	.07	.25	.26
5	.18	.22	.2	.47	.51	5	.02	.03	.06	.20	.22
6	.18	.17	.2	.46	.48	6	.02	.03	.05	.17	.19
7	.17	.16	.19	.44	.48	7	.02	.03	.05	.16	.18
8	.17	.16	.19	.43	.47	8	.02	.03	.05	.16	.17
9	.16	.14	.19	.42	.47	9	.02	.03	.04	.15	.17
20	.10	.12	.16	.30	.37	20	.02	.02	.03	.10	.12
50	.05	.07	.10	.17	.22	50	0	0	.02	.05	.07
70	.04	.06	.08	.14	.14	70	0	0	.02	.04	.05
100	.03	.04	.06	.09	.10	100	0	0	.02	.03	.04
Mean	.18	.21	.21	.41	.42	Mean	.02	.04	.06	.17	.19
S.D.	.14	.16	.11	.18	.17	S.D.	.02	.04	.03	.10	.11

two-word string is repeated quite extensively in the newspaper corpus since it accounts for .58% of all two-word strings. The table shows a consistent pattern of increased repetition from left to right at all frequency levels sampled, and for both sets of phrases. Phrases are consistently more repeated in the three learner corpora (on the right) than the two NS corpora (on the left). NS–learner differences are even greater for three-word than for two-word phrases. This is the tendency reported in de Cock et al.: not less dependence on prefabs, as Kjellmer speculated, but rather more dependence on fewer prefabs. This increased repetition is expressed visually in Figure 5 with respect to one randomly chosen phrase (item 20, in bold face in Table 3b).

FIGURE 5
Degree of repetition across corpora for two and three word strings





Interestingly, the advanced learner (TESL or 'teach') corpus appears to have more in common with the NS corpora than it has with the two ESL learner corpora. By simple *t*-tests, there is no distinction between the newspaper and advanced learner columns for two-word phrases ($p=.10$), but a strong difference between advanced and high learner corpora ($p<.001$), and the pattern is even stronger for three-word phrases. This could be taken to indicate that the phrasicons of these advanced learners are near to approximating those of native speakers.

However, such a conclusion might say more about the process of automatic corpus extraction than it does about Kjellmer's initial intuition that there is something that strikes one as odd about advanced learners' phrases.

Extracting phrases automatically from a corpus provides useful but incomplete information. For example, knowing how many phrases there are in NS and learner corpora tells us little about 'which' phrases are contained in the two corpora, or whether they are largely the same or largely different. For this, one needs either to rummage extensively through the corpora themselves, or through the extracted lists (Table 4a), or else one needs a theory about what to look for in the corpora. To take an obvious example of knowing what to look for, it is well known that English NSs make extensive use of verb-preposition combinations in both speech and writing. Is this type of phrase necessarily represented in an advanced learner corpus that has a native-like phrase count? Take, for example, verb-preposition phrases involving the preposition 'out.' A method for focusing the extraction on just these phrases is, first, to generate concordances for 'out' in each corpus, and, second, to have Wordsmith extract the recurring strings from that output and tally the frequency of each throughout the corpus. Table 5 shows the number of

TABLE 5
Phrasicons for verb-*out*

Teachers "Out"			Newspapers "Out"		
N	cluster	Freq.	N	cluster	Freq.
1	out of	37	1	out of	43
2	of the	15	2	out the	10
3	out that	11	3	of the	9
4	out to	10	4	out a	6
5	find out	9	5	out to	6
6	to find	7	6	out and	5
7	came out	5	7	out in	4
8	get out	5	8	of a	3
9	out the	5	9	of date	3
10	that the	5	10	out from	3
11	come out	4	11	out their	3
12	found out	4	12	to be	3
13	it turned	4	13	turned out	3
14	out in	4	14	broke out	2
15	out loud	4	15	come out	2
16	out with	4	16	coming out	2
17	teachers out	4	17	figure out	2
18	that I	4	18	find out	2
19	to be	4	19	found out	2
20	turned out	4	20	get out	2
21	a big	3	21	grew out	2
			22	holding out	2
Ratio		4:31 (7.8)			8:21 (2.5)

repeated strings in the concordances for the preposition ‘out’ from the TESL (NNS) and newspaper (NS) corpora (which as already noted have equal raw repeated string counts), with verb-preposition items in bold face.

Scanning this collection of verb-‘out’ phrases, one finds a pattern reminiscent of the one that was observed just above. As with phrases in general, these advanced learners clearly do use ‘out’-phrases, but fewer of them and with more repetition. Six verb-preposition strings occur in the learner corpus, but two of them are just past-present forms of the same phrase (‘come out’ and ‘find out’) so really there are only four. These four phrase types account for 31 occurrences, an average of just under eight repetitions per item. In the newspaper corpus, there are 10 verb-preposition occurrences, with two of these repeated (again ‘come out’ and ‘find out’), so that eight phrase types count for 21 occurrences or an average of 2.5 repetitions per item. To summarize, there are half as many ‘out’-phrases in the advanced learner corpus, but these are repeated three times more often. Similar results have been obtained with several other verb-preposition combinations [i]. So the impending nativeness of these advanced learner phrasicons, as suggested in Table 4b, seems less clear when specific comparisons are targeted. Obviously, there are other types of phrases that could be examined and it is not guaranteed that every examination would produce the same pattern. However, several contrastive LC studies of other phrase types have produced similar or complementary findings. Milton (in Granger, p. 189) looked at the use of discourse structuring phrases in Hong Kong academic learners’ writing and provides a interesting table of phrases overused and underused by learners in Hong Kong – overuse and underuse being two sides of the same coin.

In conclusion, the pattern is the same for phrases as it was for basic vocabulary in the replication of Ringbom: fewer items repeated more. The Québec corpus essentially confirms and strengthens the finding of de Cock et al., that advanced learners do indeed use precasts, just fewer of them repeated more frequently, as can be demonstrated by the two-step process of automatic extraction followed by targeted extraction. And yet the phrasal verb addition to the portrait rescues, to some extent, Kjellmer's idea about the scarcity of precasts. If the apparent vagueness and non-nativeness of advanced learner language cannot be put down to a simple lack of phrases per se, it may still have something to do with lack of phrase diversity, appropriateness, and nuance. Once again, this is an empirical question for which the LC study merely provides a basis to proceed. An empirical study to follow from Kjellmer, de Cock, and the present Québec replication might be to rate a set of advanced learner

texts for nativeness, or vagueness, and then compare these ratings to phrase counts and compositions. We would probably find a significant negative correlation between ratings and repetitions, but this should not be taken for granted.

Pedagogy

It appears that advanced learners either do not discover, or discover only slowly, the full phrasicon of English that is implicitly known to any NS. This is further support for the finding of Bahns and Eldaw (1993) that collocation problems remain in advanced learner English after most other problems have been resolved. Two teaching implications seem clear:

1 A new word should be taught (met, discovered, noted down, remembered) not in isolation but in the natural 'company it keeps' (Sinclair, 1991, citing Firth, 1957)—i.e., in the company of other words. One way to learn words in this manner is through natural exposure; another is with the aid of a computer concordance program linked to an appropriate corpus where every word can be met in the context of numerous and varied authentic examples and its collocations sorted by frequency [a4].
2 Some prefabs or multi-word items should probably be specifically taught. Which ones? A problem with phrases as opposed to mere words is that phrases are subject to combinatorial explosion, so that a wordlist of 2000 basic items can generate a phrase list of many more MWIs. Some selection principle would seem to be necessary, presumably frequency of occurrence either generally or within relevant domains. This point deserves elaboration.

Collecting a syllabus of MWIs is not straightforward. The frequency of individual words cannot be relied on as indicating the frequency of an MWI, as for example in the phrase 'cast against type,' where all the words are common (first 2500) items, while the phrase *qua* phrase is relatively rare (four times in the 100 million word BNC written corpus as against 757 for 'as it were.') To get round this problem, an automatic extraction program like Tuples or Wordsmith is needed that can identify the phrases that appear most commonly in the language, and – with the aid of an LC – that do not appear frequently in learner interlanguage. However, as seen above in a different context, a problem with using automatic extraction to identify a syllabus of prefabs is that many probably uninteresting and certainly unteachable phrases would appear

in the output, such as 'in the big' in the original study. While possibly frequent, such phrases are clearly not sense units and would not be useful teaching items.

A way to limit the candidate prefabs to ones that are both frequent and unitary might be as follows. Several lists of potentially interesting prefabs are available for English in both spoken and written modalities. Pawley and Syder (1983) offer a long list of speech prefabs collected on an intuitive basis, but without frequency information to signal their relative importance to learners. Milton (1998) has collected numerous prefabs from NS writing in different genres and worked out their relative importance to his Hong Kong learners, which of course might not be identical for Québec or other learners. To develop a syllabus of L1-specific prefabs, it would be a simple if tedious task to hand-enter one of these collections into a concordance program attached to a large same-modality corpus, noting the frequency of each phrase in the language at large, then do the same with a learner corpus, and subtract one list from the other. The prefabs that occur frequently in native English but infrequently in the English of one's learners would be a first guess at a syllabus of the English phrasicon.

This work could be undertaken using the resources of the Lexical Tutor. A preliminary examination of six of Milton's prefabs appears in Table 6. The numbers of occurrences for several prefabs are simply listed for each of three corpora in the second, third, and fourth columns and for the TESL (advanced) learner corpus in the fifth column. The figures in parentheses represent occurrences per 100,000 words for sake of comparison. The final column represents a syllabus inclusion decision taken on the basis that phrases should be explicitly taught if they appear more than .5 times per 100,000 words in one or more same-modality NS corpus and less than .5 times in the LC.

TABLE 6
Identifying a syllabus of prefabs for written discourse

Multi-word Item	Brown corpus 1 million wds	BNC written 1 million	BNC spoken 1 million	Learner advanced 75,000	Syllabus?
On the other hand	49 (4.9)	37 (3.7)	15 (1.5)	10 (13)	no
It can be seen	5 (.5)	7 (.7)	0	0	yes
in other words	22 (2.2)	22 (2.2)	40 (4)	6 (7.8)	no
this is not to say	3 (.3)	1 (.1)	0	0	no
for example	161 (16.1)	154 (14.4)	76 (7.6)	12 (16)	no
it is clear	15 (1.5)	13 (1.3)	0	0	yes

The criteria of .5 is just a first guess, but it seems clear that 'it is clear' is worth drawing to Québec learners' attention, while, on the other hand, 'on the other hand' is not. How prefabs should be taught, once identified, is another matter. Corpus searches are one idea; full Internet searches entering phrases into a search engine such as Google [j] is another. Groups of learners could tackle stretches of the target phrase syllabus and contribute their search results to a collaborative on-line database (such as [a6]).

Study 3

S. Petch-Tyson, *Writer/Reader Visibility in EFL Written Discourse* (pp. 107–118).

Summary

In the studies reviewed thus far, the focus has moved from words to phrases, and now it moves to discourse. The common observation tested in Petch-Tyson's study is that advanced writing, even when mainly error-free, tends to be restricted to a non-native like range of rhetorical genres. The genres investigated are spoken language and written language, orality and literacy, as distinguished by 'the degree to which interpersonal involvement or message content carries the signaling load' (Tannen, 1982, p.3, qtd. in Petch-Tyson, p. 107). In spoken language, interpersonal involvement tends to carry the signal, while in written language the signal is carried by message content. While Petch-Tyson is clearly talking the end points of a continuum that may rarely exist in the state of natural communication, it nonetheless seems clear that in NS expository or argumentative writing, reader and writer both are low-profiled as a way of emphasizing facts and issues. Does the writing of advanced learners tends to be 'talk written down,' whatever the context, situation, or genre restriction?

The background to this study is the extensive work on defining the characteristics of written texts, i.e., on defining the differences between speech and writing (Cummins, 1979; Olson, 1977; Ong, 1982; Tannen, 1982). A central idea in this work is that while spoken conversation can rely heavily on shared physical context, clarification requests, immediate confirmations, and the like, a written text must be explicit, self-contained, and comprehensible to readers far removed in space, time, and even culture from the initial context of writing. In Cummins' (1979) classic formulation, many second language learners achieve control over BICS (basic interpersonal communication skills) while fewer achieve

CALP (cognitive and academic language proficiency), the latter being strongly tied up with context reduction and the achievement of academic literacy.

Petch-Tyson's study is designated an exploration rather than a test of a hypothesis. Her purpose was to compare NS and NNS corpora for the presence and extent of spoken language characteristics. Her method was to draw up a list of features signaling reader / writer visibility, including use of first and second person pronouns ('I,' 'you,' and their variants), references to writers' mental states and processes ('think,' 'feel,' 'believe'), conversational monitoring of information flow ('you know,' 'I mean'), and others from the genre research 'that best lent themselves to automatic retrieval' (p. 110). These were then tested for presence and extent in NS and NNS corpora of equal size and task type (an argumentative essay for a university course).

The analysis found that advanced learners of four European nationalities employed from two to four times the number of spoken language features that equivalent American NSs did for an equivalent writing task, especially with regard to first and second person pronouns. Table 7 shows the extent of these pronouns as a percentage of all lexical items in each corpus for (American) NSs and for EFL learners of four European nationalities (calculated from Petch-Tyson's results which give findings as occurrences per 50,000 words).

In other words, the evidence suggests that advanced learner writing indeed resembles 'talk written down' (perhaps no surprise after 20 years' emphasis on spoken interaction in the language classroom).

Replication

The Québec advanced LC is comparable to Petch-Tyson's corpus in size and genre. It comprises short essays answering the question 'What can be done to improve the quality of English that is taught in Québec schools?' The question was formulated to elicit an expository text: it has

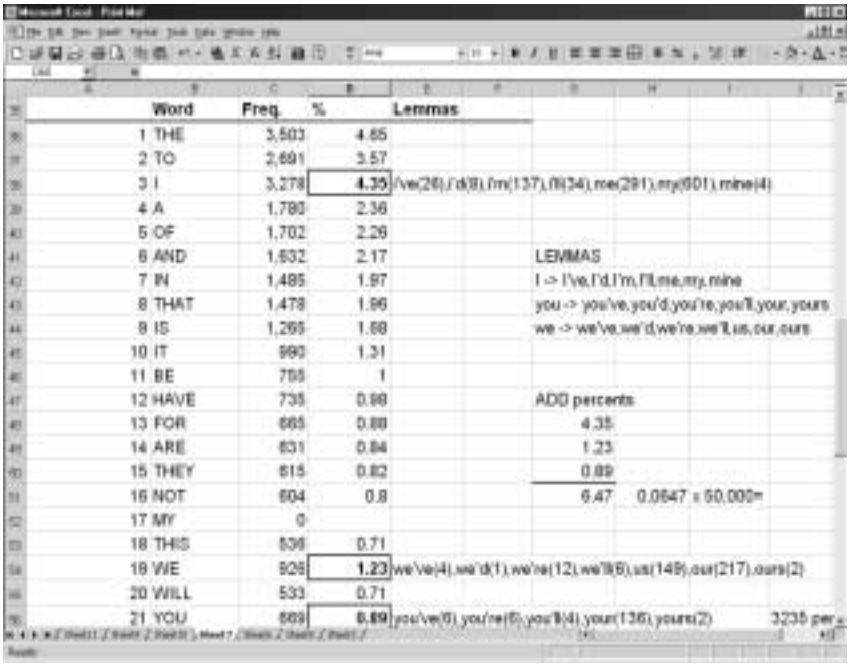
TABLE 7
First and second person pronouns in NS and advanced learner academic writing

Nationality	Occurrences per 50,000	Percentage of words in corpus
US (native)	449	0.89%
French	1,202	2.04%
Dutch	1,195	2.39%
Finnish	1,531	3.06%
Swedish	1,998	3.99%

no reference to ‘you,’ no request for a personal opinion, and the verb is in the passive voice with no obvious subject. While an opinion response would not be impossible or even out of place, one might have expected at least some proportion of the learners to adopt a detached viewpoint on the question (‘The government should ...’ or ‘Money must be spent,’ and so on).

The replication of Petch-Tyson's study will be limited to searching the Québec LC for all first and second person pronouns, the feature most indicative of reader-writer visibility in the original study. To facilitate automatic extraction, Wordsmith was used to break the LC into a frequency list, and then the program's user-defined lemmatizer-grouped pronouns of related morphology or meaning. For example, in Figure 6 the pronouns ‘I,’ ‘me,’ ‘my,’ ‘myself,’ and ‘mine’ are grouped as instances of ‘I–‘ which, thus aggrandized, accounts for 4.35% of corpus items. (Notice that ‘my’ retains its original position but has been reduced to zero coverage in the lemmatization procedure.)

FIGURE 6
List to lemma with Wordsmith



	Word	Freq.	%	Lemmas
1	THE	3,503	4.65	
2	TO	2,691	3.57	
3	I	3,278	**4.35**	/ve(26),/'d(8),I'm(137),/ll(34),me(291),my(601),mine(4)
4	A	1,780	2.36	
5	OF	1,702	2.26	
6	AND	1,632	2.17	
7	IN	1,485	1.97	
8	THAT	1,478	1.96	
9	IS	1,265	1.68	
10	IT	990	1.31	
11	BE	755	1	
12	HAVE	735	0.98	
13	FOR	665	0.88	
14	ARE	631	0.84	
15	THEY	615	0.82	
16	NOT	604	0.8	
17	MY	0		
18	THIS	536	0.71	
19	WE	926	**1.23**	we've(4),we'd(1),we're(12),we'll(6),us(149),our(217),ours(2)
20	WILL	533	0.71	
21	YOU	669	**0.89**	you've(6),you're(6),you'll(4),your(136),yours(2)

LEMMAS
I -> I've,I'd,I'm,I'll,me,my,mine
you -> you've,you'd,you're,you'll,your,yours
we -> we've,we'd,we're,we'll,us,our,ours

ADD percents
4.35
1.23
0.89
6.47 0.0647 x 50,000=

3235 per

Similar operations were performed for ‘we’ and ‘you,’ and the sums worked out: a total of 6.47% of the words in the advanced learner corpus are visible pronominal references to the writer or reader, compared to 0.89% in the NS corpus and only 2.04% in the European French LC. This would suggest that interpersonal involvement, as opposed to message content, is carrying most of the signaling load for these Québec learners. It would appear they are if anything more dependent than their European counterparts on a speech model of writing.

In both of the previous replications, it has been possible to balance the finding of a learner weakness with some sign that advanced learners are progressing slowly toward the NS norm. This was done by comparing the advanced TESL corpus with the ESL corpus. Unfortunately, such a comparison is not possible here since the stimulus topic of the ESL corpus (What difference would it make to you if your English were better?) invites a personal reflection. It is extremely important in this type of research to plan for corpus comparability.

Also, as in the previous studies, it is worth repeating that any counting-up studies provide merely correlational information and require the follow-up of other kinds of investigations. In the present instance, we find a lot of personal pronouns in learner writing, and NS instructors sense an undue degree of personal involvement in learner writing, but the causal link remains suggested rather than established. An empirical experiment that would follow smoothly here would have instructors categorize learner compositions by degree of perceived reader-writer visibility and then determine whether these judgments matched pronoun or related count-ups to any significant extent.

Pedagogy

If we can assume that advanced learners are reading widely in a variety of academic and other text types, then it seems clear that reading alone is not enough to expose the main features that distinguish the various genres, such as the low writer-reader visibility that typifies an argumentative text. There is a case for mixing these learners' reading with focused awareness raising of the formal features of different genres, perhaps accompanied by an occasionally reversion to the outmoded practice of writing from models.

Conclusion

It appears that even advanced learners are unlikely to discover very quickly on their own all of the relevant features of a second language

that make it native-like. This is not surprising since, as the original motivation for these LC studies makes clear, even experienced language teachers have been unable to pin down what is missing from advanced learner language. The features are perhaps too diffuse and numerous, and that is why methods and instrumentation that go beyond observation are needed to disclose them.

The distinctions between native and learner English that have been tested here are guesses that have often been put forward by linguists, educators, and teachers, but until now there has been no systematic way to test or refine these guesses. Corpus analysis makes this possible, with promising results which appear to be rather robust. And yet as has been noted for each of the studies, contrastive learner corpus analysis is most useful as a step between intuition and hypothesis. LC analysis can establish the plausibility, for example, that overuse of words and phrases is a source of perceived non-nativeness in advanced learner writing, but only a controlled experiment can prove it. In the three-part research agenda proposed for SLA by Long (1983) and others, where description is followed by correlation and culminates in controlled experiment, learner corpus analysis will probably be important in the second or correlational phase. This phase is too often omitted in our research area, possibly because of the prestige afforded to the one-off controlled study. However, controlled studies are most valuable if the plausibility of the hypothesis has been well established beforehand. In other words, corpus analysis presupposes a research agenda where knowledge is built up over a series of complementary investigations of phenomena.

This paper has argued that a research agenda which includes contrastive learner corpus analysis can shed light on the nature of advanced interlanguage. In the introduction, it was proposed that instruction can do more for these learners than just give them 'lots of practice.' The evidence presented here suggests that advanced learners are not defective native speakers cleaning up a smattering of random errors, but rather learners working through identifiable acquisition sequences. The sequences are not the *-ing* endings and third person *-s* we are familiar with, but involve more the areas of lexical expansion, genre diversification, and others yet to be identified. That these sequences are systematic and more or less universal is suggested by the similarity of findings across several first languages and now across the Atlantic. Learner corpus analysis should open new vistas for North American researchers and learners alike.

The first step in this agenda is, of course, corpus building, and here we are quite far behind our European colleagues. For readers interested in participating in this interesting and useful line of investigation,

particularly with regard to collaborative corpus building, the web site of the ICLE (International Corpus of Learner English) [k] at the Université catholique de Louvain, Belgium, is a good place to start.

Tom Cobb is a professor of TESL in the Dépt. de linguistique et de didactique des langues at the Université du Québec à Montréal, where he teaches courses involving the uses of computing in language teaching and learning. He has taught ESL in Canada as well as Saudi Arabia, Oman, and Hong Kong. He holds a PhD in educational technology from Concordia University.

Notes

1 Of course learner corpus analysis is not the only current approach to a better characterization of advanced IL. Other branches of applied linguistics are also working on this problem. The recent studies in fluency (CMLA 2001) and automaticity (Segalowitz, 2000) are related attempts to locate advanced learners within a multifaceted developmental profile. The fluency studies show what learners can do with their L2; the corpus studies show what they know about the L2—and, sometimes even more interesting, what they do *not* know about it.

2 As mentioned, it has proven difficult to collect longitudinal data from advanced learners. It is therefore worth pointing out regarding Table 1 that corpus studies typically construct tendencies on the basis of cross-sectional data, treating the cross sections as sequential despite the fact that they were not produced by the same people (this was also done in some of the early child IL studies). This practice will be adopted here.

3 Such remarkable consistencies are a feature of corpus studies, whether in linguistics or applied linguistic studies, and are one of the satisfactions of work in this area.

References

Alderson, J. (1984). Reading in a foreign language: A reading problem or a language problem? In J.A. Alderson & A.H. Urquhart (Eds.), *Reading in a Foreign Language* (pp. 1–27). London: Longman.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System 21,* 101–114.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge, UK Cambridge University Press.

Bloom, L., Hood, L. & Lightbown, P. (1974). Imitation in language development: If, when, & why. *Cognitive Psychology 6,* 380–420.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Cobb, T. (1997). *From concord to lexicon: Development and test of a corpus-based lexical tutor.* Unpublished doctoral dissertation. Concordia University, Montreal.

Cobb, T. (1995). *Imported tests: Analyzing the task.* Paper presented at TESOL (Arabia). Al-Ain, United Arab Emirates, March.

Cobb, T. (2001). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review*, 57 (2), 295–324.

Coxhead, A. (2001). A new academic word list. *TESOL Quarterly, 34* (2), 213–238.

de Cock, S., Granger, S, Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In Granger (Ed.), *Learner English on computer* (pp. 67–79). London: Longman.

Cummins, J. (1979). Cognitive / academic language proficiency, linguistic interdependence, the optimal age question, & some other matters. *Working Papers on Bilingualism 18,* 197–205.

Dulay, H., & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning 24,* 37–53.

Firth, J.R. (1957) A synopsis of linguistic theory, 1930–1955. In Palmer, F.R. (Ed.) (1968) Selected papers of J.R. Firth 1952–9. Harlow, UK: Longman.

Granger, S., Ed. (1998). *Learner English on computer*. London: Longman.

Kjellmer, G. (1991). A mint of phrases. In Aijmer, K., & Altenberg, B. (Eds.), *English corpus linguistics* (111–127). London: Longman.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16,* 307–322.

Long, M. (1983). Inside the 'black box': Methodological issues in classroom research in language learning. In H.W. Seliger & M. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 104–123). Rowley, MA: Newbury House.

Meara, P. (1993). Tintin and the world service: A look at lexical environments. *IATEFL: Annual Conference Report,* 32–37.

McEnery, T. & Wilson, A. (1996). *Corpus linguistics.* Edinburgh, UK: Edinburgh University Press.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on computer* (pp. 196–198). London: Longman.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, & pedagogy* (pp. 40–63). Cambridge, UK: Cambridge University Press.

Nation, P. (1990). *Teaching and learning vocabulary*. Boston: Heinle.

Olson, D. (1977). From utterance to text: The bias of language in speech & writing. *Harvard Educational Review 47,* 257–281.

Ong, W. (1982). *Orality and literacy: The technologizing of the word.* New York: Routledge.

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

Petch-Tyson, S. Writer / reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107–118). London: Longman.

Pienemann, M. (1999). *Language processing and second language development.* Amsterdam: John Benjamins.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In Granger (Ed.), *Learner English on computer* (pp. 41–52). London: Longman.

Segalowitz, N. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on fluency*. Ann Arbor, MI: University of Michigan Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics 10,* 209–31.

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

Stubbs, M. (1996). *Text and corpus analysis.* Oxford: Blackwell.

Tannen, D. (1982). The oral / literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 1–17). Norwood NJ: Ablex.

Wray, A. (2001). *Formulaic language and the lexicon.* Cambridge, UK: Cambridge University Press.

Appendix: Web sites

[a] Cobb, T. *Compleat Lexical Tutor*
http: / / 132.208.224.131 /

[a1] Web Concordancer
http: / / 132.208.224.131 / Concord.htm

[a2] Web VocabProfile
http: / / www.er.uqam.ca / nobel / r21270 / cgi-bin / webfreqs / web_vp.cgi

[a3] Web Frequency Indexer
http: / / www.er.uqam.ca / nobel / r21270 / textools / web_freqs.cgi

[a4] List Driven Learning
http: / / 132.208.224.131 / ListLearn /

[a5] British National Corpus frequency lists for words of 800-plus occurrences in 100 million

http://132.208.224.131/BNC_numerical.txt
[a6] Group Lex collaborative on-line database
http://relish.concordia.ca/esl298b/tom_php/lex.php
[b] Kucera, H. & Francis, W. *Brown Corpus Manual* on-line (1979)
http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM
[c] Mike Smith's *Wordsmith* web site, Liverpool University.
http://www.liv.ac.uk/~ms2928/wordsmith/
[d] Nation, P. LALS web site Victoria University, NZ.
http://www.vuw.ac.nz/lals/staff/paul_nation/
[e] Invisible Lighthouse Frequency page
http://www.invisiblelighthouse.com/langlab/bncfreq.html
[f] Adam Kilgarriff's collection of BNC frequency lists
http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html#bib
[f1] Sample search from 100-million-word British National Corpus
http://sara.natcorp.ox.ac.uk/lookup.html
[g] Averil Coxhead's Academic Word List page
http://www.vuw.ac.nz/lals/div1/awl/
[h] John Milton's *Wordpilot* (Hong Kong University of Science and Technology)
http://home.ust.hk/~autolang/whatis_WP.htm
[i] AAAL PowerPoint presentation of this paper
http://www.er.uqam.ca/nobel/r21270/cv/QLCorpus/QL_Corpus.htm
[j] Google Internet search engine
http://www.google.com/
[k] CECL (Centre for English Corpus Linguistics) - ICLE Web site
http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm