CHAPTER FOURTEEN

A RESOURCE WISH-LIST FOR DATA-DRIVEN LEARNING IN FRENCH

TOM COBB
UNIVERSITY OF QUÉBEC AT MONTREAL, CANADA

Introduction

The following reflection on the current and potential state of data-driven learning (DDL) opportunities for learners of French as a second language (FSL) is written from the perspective of an English as a second language (ESL) developer in computer-assisted language learning (CALL) who has adopted a DDL approach for a wide range of practical web-based ESL applications and now is under growing pressure to do the same for FSL learners. The author is not a specialist in French linguistics or corpus studies, and in what follows will be talking about the challenges of meeting practical needs and finding materials to accomplish some of the same goals in FSL as have been accomplished in an ESL context. The applications in question are housed on a website entitled *The Compleat Lexical Tutor* (www.lextutor.ca), whose tracking statistics show a steady growth in the number of FSL users, despite its only basic capacity for handling the French language. From this growth it is assumed there is a demand for DDL materials in FSL beyond what teachers and learners are finding readily available, leading to the question of what would be needed to meet this demand more adequately. The discussion will proceed from a definition of hands-on DDL, to a discussion of what can be done in French on Lextutor with minimal resources, to what would be needed for a fuller implementation.

What is DDL?

What is DDL and why does French language learning need it? The answer to the first question begins, in the manner of the best French philosophical writing, with a conceptual taxonomy.

DDL could refer to any approach to language learning with an emphasis on authentic input (see Boulton 2011 for an investigation of the term), but here it will refer to a computational approach to language learning within the input-driven paradigm. 'Data', in this view, is a particular interpretation of 'input' and basically means 'computer processed input'. Learning language through input is a very broad idea with many realisations, held together mainly by downplaying biological intuition or "innate learning algorithms" (Pinker 1994) as the primary source of language acquisition. Given enough time and exposure, input theories propose, language learners gradually respond to and reproduce the underlying lexical, grammatical, pragmatic and other patterns that are implicit in the languages they encounter, whether through unconscious habit formation (from a behaviourist or more recently an emergentist perspective), or through some element of conscious noticing (from a cognitivist or more recently language awareness perspective). DDL is an input-based approach to learning second languages (L2s) which emphasises the role of awareness, but contributes a further nuance to the many similar approaches: it assumes that naturally occurring input will reveal its patterns more slowly and obscurely than is often required in learning an L2, but can reveal them more quickly and clearly when reconfigured as data and run through certain kinds of computer programmes. Such programmes expose the patterns in data, enabling a learner to transform data into information.

The most basic of these computer programmes are frequency analysers (which determine the number of occurrences of words or phrases in texts or corpora) and concordancers (which do the same but also show a short context for each occurrence). The language patterns that these programmes can bring to a learner's awareness might include the following: a frequency programme can reveal the words or expressions that recur most often in a text, as could be discovered only slowly through reading the text, or might not be discovered at all. A concordance programme can show the most frequent neighbours or associates of particular words in a text or in a language as a whole (e.g. *drive* for 'car' and *ride* for 'bicycle'). A considerable body of research now shows that frequent associates are often the last things that learners discover in their L2s (e.g. Nesselhauf 2005) and may never be discovered at all.

Drilling down into the taxonomy, a further nuance is whether learners will use software to discover the patterns in linguistic data for themselves or take somebody else's word for such discoveries. For example, lexicographers might use software to query a large corpus of a language, determine which words are more worthy of attention than others, and then incorporate this information into a dictionary for learners, who may or may not know or care where the information came from or what its basis is. Alternatively, learners might be shown how to obtain this type of information for themselves, in which case DDL is allied with a form of discovery or constructivist learning in which proactive learners-as-linguists (or "detectives," Johns 1997) form their own questions and seek their own answers in the data of a language.

This latter definition is the principle meaning proposed in the present discussion and is summarised in Figure 1. To recapitulate, core DDL is language learning based on input, reconstrued as computer-manipulated data, as inspected by learners themselves (the bottom-right box in the diagramme).

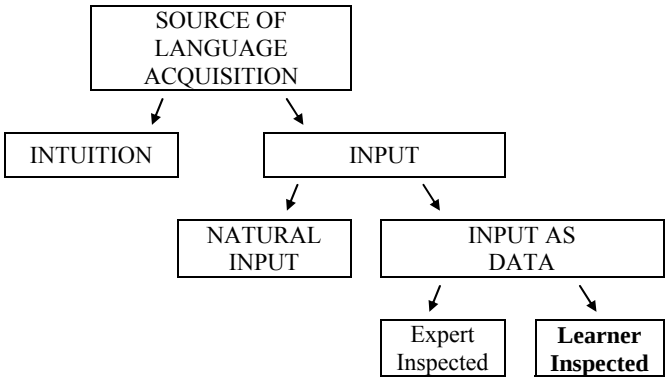


Figure 1. The place of data-driven language learning in the broader scheme

It is probably safe to say that while various aspects of DDL now pervade the language teaching industry (corpus-derived learner dictionaries, e.g. the *Longman Dictionary of Contemporary English* 2011; grammars, e.g. Biber *et al*.'s 1999 *Longman Grammar of Spoken and Written English*; and whole language courses, e.g. McCarthy *et al*.'s 2006 *Touchstone*), these are nonetheless not instances of core DDL but rather 'expert inspected data' in terms of Figure 1 (the bottom left box). While

either brand of input-as-data could be applicable to the discussion that follows, most of the examples will be from the 'learner inspected' corner.

But first a clarification: the distinction between learner and expert inspected data while useful is not absolute. It is entirely possible and quite common for linguistic data first to be inspected by an expert and then modified in some way for subsequent delivery to learners as data for them to investigate still further. In this sense the data which the learners' investigate is not entirely 'raw', and yet it is still more raw than, for example, a word definition written by lexicographers that is 'based on' corpus data. As with many types of discovery learning, it is often necessary to set up the data to allow the discovery to be made (Cobb 1999). For example, learners are unlikely to discover very many collocational patterns in a text or corpus with more than a certain percentage of unknown lexis, so the lexical level must be controlled in order for the collocational discovery to be made. This discussion will be further elaborated in a discussion on the role of frequency lists later in the chapter.

There is a legitimate question about learners inspecting data with computer programmes at all, regardless of whether the data is raw or merely 'rawish'. On one side, there is strong and longstanding support in second language acquisition research for pattern extraction by learners with minimal preconception as to what they will find (Ellis & Schmidt 1997); support for learner-as-scientist learning models in specific areas like vocabulary (Cobb 1999); and support for self-initiated, effortful learning generally (Hulstijn & Laufer 2001). On another side, there is some question about what sense learners can make of linguistic data, and at what point and in what form it should be introduced (Boulton 2010). The position adopted in this chapter is that learners can be involved in making sense of the output of linguistic computer programmes—data—from very early stages in the learning process, provided the data and the software are adapted to their needs and interests. For example, the data need not be a general corpus of a language but might be a collection of simplified stories. Concordance output need not be pages of concordance lines, but might be a short sample adapted to a particular task or game. Figure 2 depicts an ongoing word-review activity (from http://www.lextutor.ca/id), where words met previously in a story text are being recalled in an environment of several further novel instances of each word. The goal is to identify the word by selecting it from a jumble of letters with the mouse; the cue is 12 lines from a corpus of simplified stories from Oxford's Bookworm series.

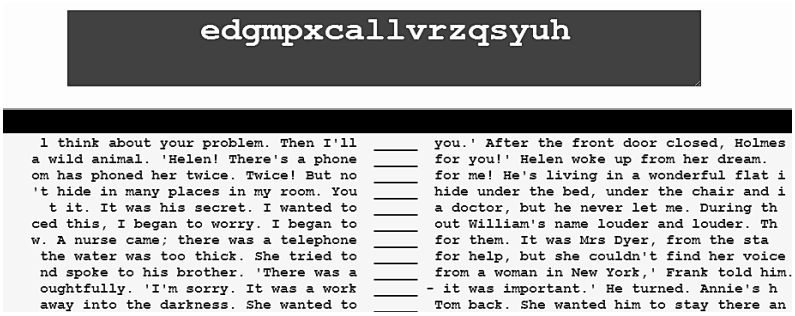


Figure 2. Data processing for beginners

The activity shown in Figure 2 and many others like it comes from the author's Lextutor website (www.lextutor.ca), which occupies an interesting place both in the DDL landscape and in the present discussion. First, Lextutor's many applications of the DDL concept are both widely used (10,000 average consultations per day worldwide) and have been developed and refined in collaboration with teachers and learners over an extended period. But Lextutor also provides one of the few sources of information about the true extent of hands-on DDL as a phenomenon. Whether and how much language learners choose to engage with data as a source of language acquisition has traditionally been hard to determine. The classic software tools for ESL like Johns' (1986) MicroConcord were popular products, but no one knew how much they were used, what corpora they were used to analyse, or whether their users were researchers, teachers, or learners. In contrast, online software like Lextutor provides usage data making it reasonably clear that much of the activity is coming from learners. For example, the concordances are largely generated in the context of a learning activities such as the one shown in Figure 2. Many of Lextutor's concordancing activities (such as MultiConc at http://www.lextutor.ca/concordancers/multi) have both a quiz and a research option, and the quiz option is chosen three times more frequently. Furthermore, the corpora chosen are typically those that would be of more interest to learners than to researchers. The most used is the 2-million word collection of simplified stories already mentioned. Web statistics are a blunt research tool, but in this case the volume of consultations per day suggests that hands-on, learner-initiated DDL is quite widespread.

So far we have seen that DDL enjoys a basis in both acquisition research and ongoing practice, but it may also have other benefits. One is that it could provide a more principled basis for the development of computer-assisted learning. What we have at the moment is a mainly

unprincipled assembly of paper-based multiple-choice questions adapted for the Internet that do not actually exploit computing power to any significant degree. Another potential benefit is face validity. It is easy to imagine that advanced language learners, such as foreign doctoral students who are studying in a second language, may tire of classroom language learning led by a modestly educated instructor well before they have actually developed the L2 competencies they need and would welcome a more 'scientific' avenue for their learning (examples of challenging independent learning activities are shown below). Another benefit is the plausible but so far uninvestigated transfer of DDL investigations to independent learning. The mental habit of querying a corpus for its patterns possibly extends to a long-term strategy of querying the L2 in the corpus of the environment with a similar objective. Yet another is that DDL explicitly focuses learner attention on the real L2 'out there' rather than on the dilute sample presented in a course book. And finally the investment already put into DDL in English should transfer fairly simply to other languages in the same language type (with a Roman alphabet) including, and indeed notably, French.

Why does French language learning need DDL?

French is a promising language for the exploitation and further development of DDL tools and principles. French corpus research is active (e.g. http://www.lexique.org), and at first sight French corpora appear to be plentiful. Interest in literature-based and prescriptivist or 'knowledge transmission' approaches to learning French appears to be declining (Boulton 2010: 547, though perhaps less in France itself than in other places where French is taught extensively, such as Canada, Great Britain, and the United States), just as interest in French for academic, professional, immigration and other specific purposes (continuing the trend of *Le français fonctionnel* in the 1970s; see http://www.le-fos.com/historique-4.htm), continues steadily to gain ground replacing the biases of earlier phases, again in some places more than others. Finally, computer-assisted learning in French as a second language, while high quality in terms of production values, and extensively used, arguably awaits an organising principle such as DDL might provide. An earlier attempt to build a systematic, corpus-based French course (Biggs & Dalwood 1976) failed to make a lasting impact on the teaching of FSL, but of course that was before computing and CALL offered anything like the possibilities they do today.

And yet the extension of DDL tools from English to French even in 2012 is not entirely straightforward. English is in many ways a privileged language for DDL purposes, with its large number of learners, its well-developed networks among practitioners, its long uncoupling from prescriptivism, literary studies, and theoretical linguistics, and the length of time corpus-based learning resources have been under development. Some of the issues involved in the adaptation of DDL to even a language as similar to English as French have become apparent through several years of running parallel French DDL services on Lextutor using only a handful of small corpora assembled by French applied linguists and donated to Lextutor. These include 1 million words from *Le Monde* in 1986, collected by Thierry Selva; a 150,000-word corpus of spoken French collected by Kate Beeching; and various literary collections found on the web. The DDL activities that can be built on these have nonetheless proven popular, claiming about 10% of Lextutor's user base or 1,000 web consultations per day. Other Lextutor routines, however, have proven impossible to adapt to French because they depend on the pedagogical adaptation of corpora in ways that are yet to be performed for French. In the rest of this chapter, these French developments and challenges for Lextutor, viewed as a test case for DDL in French, will be elaborated in more detail. A preview of the remainder of this discussion is as follows: first, straightforward transfers of DDL from English to French; second, problematic transfers; and third, a summary of the resources that seem yet to be developed in French to render a fuller DDL implementation possible.

Straightforward text and small-corpus adaptations

Straightforward French DDL adaptations will consist of a reading tool, an error-correction tool, a writing tool, a listening tool, and an analytical or language awareness tool. These have been selected from Lextutor's many routines as representative data-oriented ways of focusing on the main areas of language development (written as well as spoken, receptive as well as productive).

Hypertext: a DDL reading tool

As we have seen, corpus investigation can be adapted to learners' levels and purposes. One of these is to offer further examples for the unknown or lesser-known vocabulary in an intensive reading activity—in other words as a look-up tool, but one more effortful than a dictionary and with some claim to offer better support for learning and retention (Cobb 1997). This

activity is assembled by a learner or a teacher by simply typing or pasting the to-be-read text into a window and clicking 'Build' (see Figure 3).

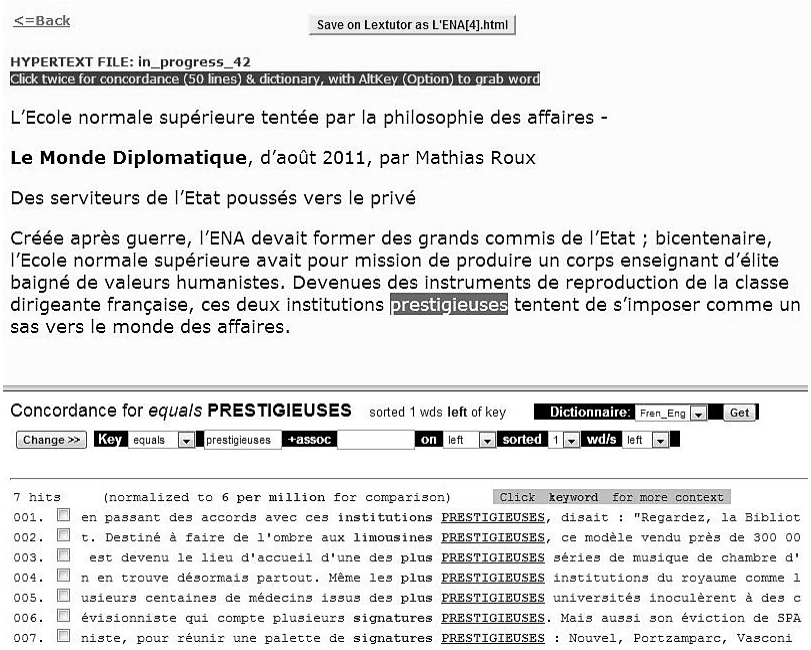


Figure 3. Hypertext Builder output

The programme output is the chosen reading text linked to a routine that generates concordances from *Le Monde* (1986) for any word double-clicked. The technology has been designed to facilitate a quick look-up with only a minimal departure from the text itself. In this case, *prestigieuses* is the word of interest, and the seven extra occurrences offer an expanded contextual space for either inferring or confirming the word's meaning. The pattern to be extracted from the data in the present case is thus the base meaning of the word underlying the several instances. The concordance also indicates that *institution* and *signature* seem to accompany the word frequently and so are potential collocates. A longer context can be generated for any line by clicking the keyword, and a dictionary is also available from the concordance output. Learning effectiveness validation for use of this software can be found in Cobb (2009). It should be noted that this activity is available for any machine-

readable French text whatsoever, although it is only appropriate for texts with a challenge level that is roughly in line with that of *Le Monde*. A French corpus of simpler texts suited to less proficient learners (like the English graded readers corpus shown in Figure 2) seems not to be available.

Concordance Feedback

A perhaps more typical DDL activity is one that invites learners to extract formal, rather than semantic, patterns from linguistic data. The Lextutor routine Concordance Feedback is one approach to doing this, in the context of errors that learners have made in their writing. The technology (elaborated in Gaskell & Cobb 2004) involves a teacher designing a hyperlinked concordance request that highlights, through several examples, the appropriate way of doing what the writer has tried to do in his or her sentence but in a non-standard manner. Normally this concordance link would be inserted directly into the learner's text, but in Figure 4 the link (here designated 'CONC') appears in a tutorial programme that prepares learners to use this type of feedback.

The element that is wrong, in this case missing, in the sentence *Je peux répondre cette question* ('I can answer this question'), is clearly indicated in the concordance output. Learners use the concordance information to perceive the pattern underlying the instances (in other words to note the missing *à*), make their correction, and then click 'Check' to see if they were right. Since a teacher has set up the concordance to make the point about the error, Concordance Feedback is an instance of learners working on data that was first worked on by a teacher or other expert. Nevertheless, as Gaskell and Cobb (2004) point out, the end goal is for the learner to work with concordance data independently to fix their own errors – or ideally not make them in the first place.

The French part of this application at this time is just a few problem sets within the English application, but in any case the tutorial is only a trainer to prepare learners to profit from concordance feedback links extracted by a teacher and inserted into their own on-line submissions as a means to understanding their own errors. The link extraction procedure is fully functional in both languages, albeit with a limited range of corpora in French (visit the French link extractor at the bottom of the page at http://lextutor.ca/concordancers/concord_f.html).

Figure 4. Concordance-based error correction

ConcordWriter: A DDL writing tool

Another learning tool that works reasonably well with the *Le Monde* corpus is ConcordWriter (at http://conc.lextutor.ca/concord_writer), a writing tool that encourages L2 writers to consult a concordance independently as they produce a piece of writing. The reasoning here is that when writing, L2 learners' attention is typically held at the level of form (word choice, spelling, collocation) such that idea generation and development become secondary. And yet, virtually every question about linguistic form that a learner is likely to have can be quickly answered in any corpus of even modest size, freeing up resources for focus on meaning. ConcordWriter allows writers to query the *Le Monde* corpus without leaving the page they are working on.

The queries are thus independent, rather than teacher constructed, and yet the principle of adapting software to learners' needs and interests is maintained. For example, shortcuts are built into the programme based on teachers' experience with the type of problems L2 writers are known to have. As mentioned, one such problem is collocation—what word is needed next in a sequence, or more broadly what type of word (singular, plural, etc.). Thus ConcordWriter allows 'starts-with' searches for the last 2, 3, or 4 words already written. This type of search will often give good information about the type of word that is needed next, or sometimes even the exact word. For example, in Figure 5, the writer has launched a French expression based on the English expression 'one of the best + noun,' and has realised she does not know if the forthcoming noun is singular or plural. Her text at the point shown ends with *un des meilleurs* ('one of the best'). By clicking 'Text End-last 3,' (option 3 in the line beginning 'Search modes'), she generates a starts-with concordance for her own three final words containing ample evidence from other writers' texts that the forthcoming noun must be plural. It also reminds her that the adjective *meilleur*s must also be plural since the noun is plural, and raises the possibility that *un* might need changing to *l'un*. The goal is to facilitate independent corpus consultation without taking writers very far away from their texts – indeed, while integrating the learners' texts with the corpus. Learning effectiveness validation for use of this software is thus far anecdotal.

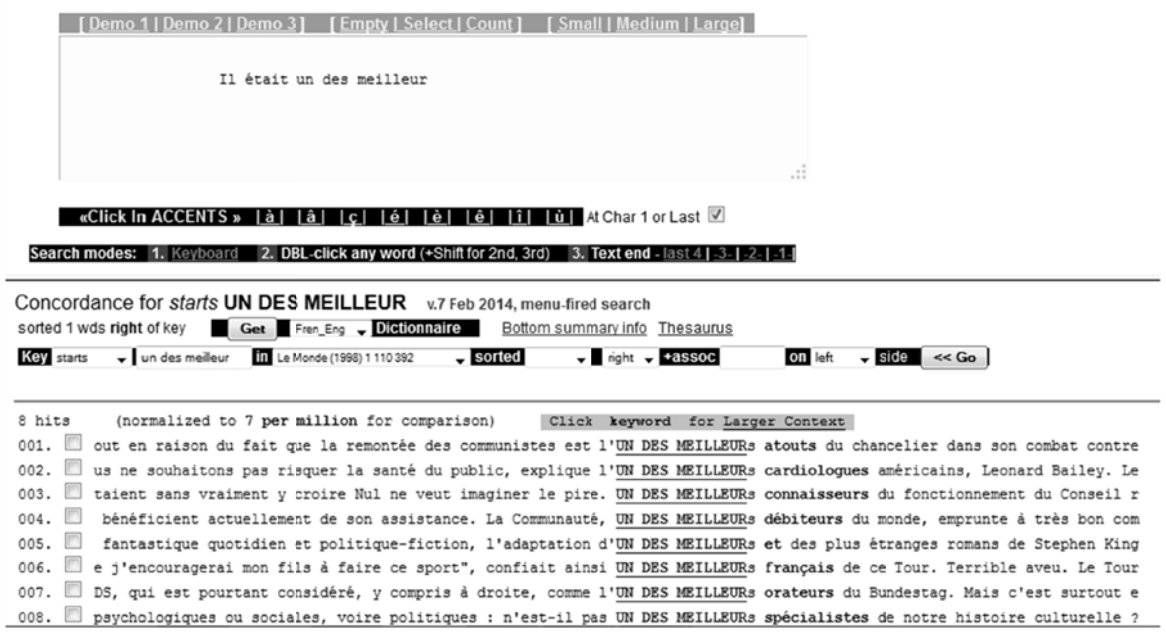


Figure 5. A ConcordWriter session in progress

Dictator: a DDL listening tool

The 'data' of data-driven learning is most commonly thought of as concordance lines, but computed language data is not limited to these formats. Another computational form of language data is speech generated from text by a computer programme, text-to-speech (TTS). TTS was considered by many until recently as 'computer' in the sense of mechanically harsh and non-human sounding, but TTS has improved markedly with significant recent investments. Unfortunately, it has also become correspondingly less accessible to CALL developers, but there are workarounds. The current workaround on Lextutor involves third-party use of Google Translation's excellent speech rendition.

The purpose of Dictator is to give learners as much listening input as they wish, plus an opportunity to listen repeatedly and carefully, attempt to write out the text of what they have heard, and then receive feedback as to their comprehension, and incidentally on their understanding of grammatical, morphological, and orthographical patterns. Either teacher or learner can type or paste words, phrases, random sentences, or sequential sentences into the input space shown in Figure 6a, choosing the desired parameters for the activity (random or sequential, etc.). In this case it is a short narrative French text (from the multilingual web site http://www.lonweb.org), so the randomisation is set to 'No' in order to retain the story line. The input sentences are separated by vertical bars.

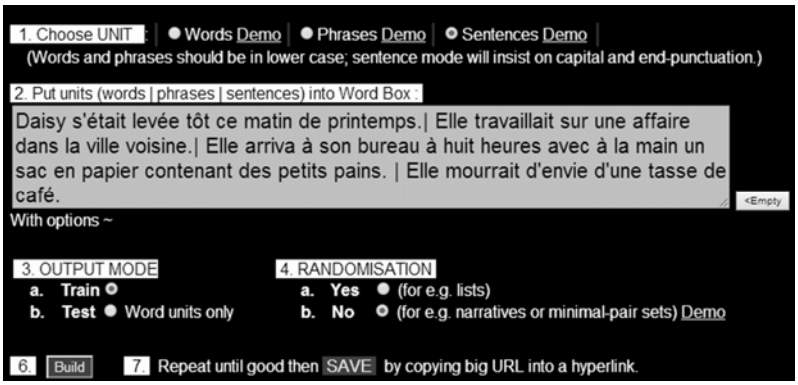


Figure 6a. Dictator input

Clicking 'Build' creates the activity shown in Figure 6b. For each input sentence, there is a button to play the sentence, a space to write the sentence, and a 'Check' button to tell the programme's Guidespell sentence-comparison algorithm to compare what the learner has written to the original sentence and indicate either a match or an enumeration of the differences. The learner working in Figure 6b has correctly reproduced the first sentence, *Daisy s'était levée tôt ce matin de printemps* ('Daisy woke early on this spring day'), except for the accents (accents can be input from a menu if unavailable from the keyboard) and the spelling of *printemps*.

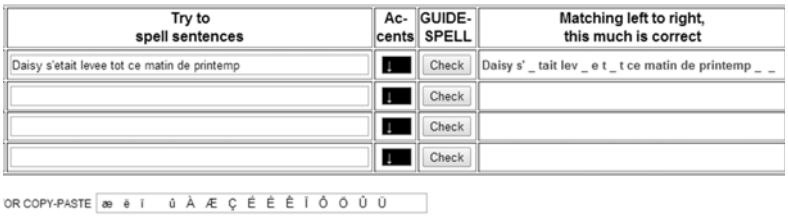


Figure 6b. Dictator output – activity under way

Range: a direct DDL awareness-raising tool

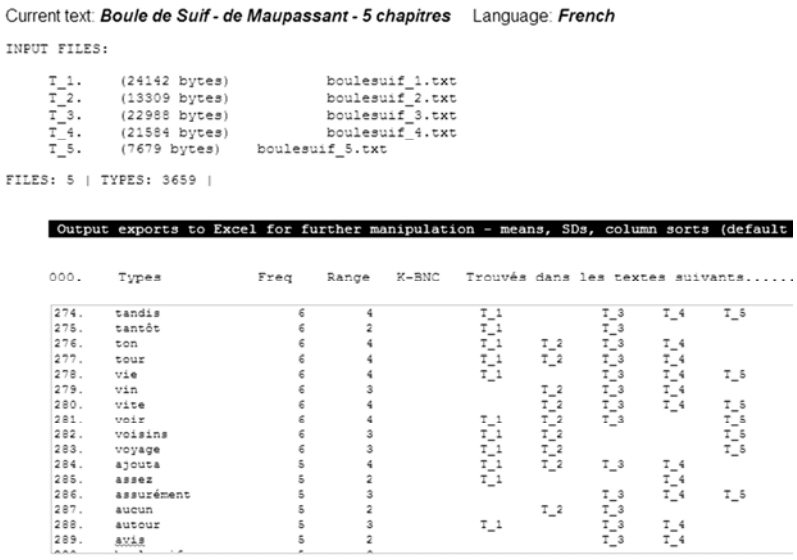
Many of Lextutor's DDL routines can be used as awareness-raising activities, and indeed many of its practitioners see this as the essence of the approach (in contrast to the more task integrated approach shown for example in ConcordWriter). Of the many possible questions about language patterns that can be answered with text analysis tools, the example of feature distribution (or 'range') will be illustrated here, since this can function reasonably well on Lextutor in French. The online version of the programme Range (at http://www.lextutor.ca/range_corpus), as well as its offline progenitor by Nation and Heatley (2002), shows the distribution of a linguistic feature in different texts or corpora or in different parts of a corpus. The English part of Range includes several pre-established types of comparisons. For example, the word *analysis* (or the whole family of related words if wildcarded as *analy**) can be traced to the parts of the BROWN corpus that it lives in: 145 occurrences in all, 12 in the press sub-division, 4 in fiction, and 129 in academic. Or it can compare a term between the written and spoken one million-word samplers of the British National Corpus (2005), showing for example that *analy** occurs only 21 times in the million spoken words but 210 times in the written

million, a point of some interest to anyone learning to use English. Clearly, the *analysis* family is more the stuff of books than of everyday speech.

A similar comparison option for French has been set up using a 150,000-word corpus assembled by Kate Beeching (available online at http://www.llas.ac.uk/resources/mb/80). The Range software allows users to make comparisons between this speech corpus and an equal sized portion of Selva's *Le Monde* written corpus. Running *analy** against these two corpora provides just two occurrences in the spoken and 27 in the written, a ratio of 1:13.5. One can further compare this finding to the English distribution in equal size corpora for this same term, 1:10, which is remarkably similar to the French proportions in this case. The argument for encouraging learners to engage in this type of meta-reflection about language and languages is found in the language awareness literature.

Another version of the Range idea runs on texts provided by the user rather than corpora provided by the system. The user loads in several text files (say the chapters of a book), and the programme shows the distributions of every word form across the series. In Figure 7, a user has loaded all five chapters of de Maupassant's *Boule de Suif*, with the output showing all the novella's words in order of frequency and range (or distribution). Points of interest in the output would be that *dames* ('ladies') appears in four chapters out of five, and either *Prussien* or *Prussiens* in all five chapters, and also *neige* ('snow') in all five, as only makes sense as this is a story of Prussian soldiers relating with French women in a winter landscape. This type of information could be useful in writing a summary of the story – or more likely, in an L2 context, of helping decide which words are worth looking up as a function of whether or not they will soon be needed again.

The interface shown in Figure 7 has clearly been designed for English (for example, the 'K-BNC' column is empty, referring to rank in the BNC frequency list). But all of Lextutor's routines are gradually being expanded to incorporate any Roman alphabet language, and following this French and other interfaces will be added.



Current text: ***Boule de Suif - de Maupassant - 5 chapitres*** Language: ***French***

INPUT FILES:

T_1.	(24142 bytes)	boulesuif_1.txt
T_2.	(13309 bytes)	boulesuif_2.txt
T_3.	(22988 bytes)	boulesuif_3.txt
T_4.	(21584 bytes)	boulesuif_4.txt
T_5.	(7679 bytes)	boulesuif_5.txt

FILES: 5 | TYPES: 3659 |

Output exports to Excel for further manipulation - means, SDs, column sorts (default

000.	Types	Freq	Range	K-BNC	Trouvés dans les textes suivants.....				
274.	tandis	6	4		T_1		T_3	T_4	T_5
275.	tantôt	6	2		T_1		T_3		
276.	ton	6	4		T_1	T_2	T_3	T_4	
277.	tour	6	4		T_1	T_2	T_3	T_4	
278.	vie	6	4		T_1		T_3	T_4	T_5
279.	vin	6	3			T_2	T_3	T_4	
280.	vite	6	4			T_2	T_3	T_4	T_5
281.	voir	6	4		T_1	T_2	T_3		T_5
282.	voisins	6	3		T_1	T_2			T_5
283.	voyage	6	3		T_1	T_2			T_5
284.	ajouta	5	4		T_1	T_2	T_3	T_4	
285.	assez	5	2		T_1			T_4	
286.	assurément	5	3				T_3	T_4	T_5
287.	aucun	5	2			T_2	T_3		
288.	autour	5	3		T_1		T_3	T_4	
289.	avis	5	2				T_3	T_4	

Figure 7. Ranges of words across the chapters of a text

Group Lex: an ongoing application of data-driven learning in French

A current project in a *francisation* programme at a Montreal CEGEP (*Collège d'enseignement général et professionnel*, an institution midway between high school and university) pulls together many of the threads from above in an ongoing adaptation of Lextutoring to the needs of French learners. Francisation is a programme for immigrants or others who for any of several reasons need to quickly develop their ability to use and understand French. Building up a lexicon of basic French words (3,000 word families at an absolute minimum, as estimated by Nation 2006) is a key goal in this process, but yet there is no agreed method to achieve it.

One approach to developing a large-scale lexical syllabus is to have learners contribute their own words to a networked lexical database on Lextutor, known as a Group Lex, where it can be further processed, reorganised, quizzed, tested, and extended to novel contexts. The provenance of these words can be course materials, the chapters of a particular book, the linguistic environment generally, or even assigned items, as decided by a teacher or programme developer. To operate the

programme, learners enter a word or phrase, example sentence, part of speech, and meaning, and all this information appears as a line in a sortable database linked to further resources. The resulting accumulation of lexical information (Figure 8) is essentially a collaborative dictionary, except that it has been built from the data up rather than the definition down, in a procedure whose full rationale and two empirical work-outs can be found in Cobb (1999) and Horst *et al*. (2005). Over 100 such Group Lexes for ESL learners have been created since 2005, and many are in constant use. Learners seem to appreciate Group Lex because they can generate interactive gap-fill quizzes for different sets of words (Figure 9), including those entered by a particular friend or acquaintance, and because they can hear the word, the example, or the definition that they or a classmate have entered. Teachers seem to appreciate being able to check whether students have done their work and to generate daily or weekly paper quizzes. Researchers seem to appreciate the host of researchable questions this software generates (such as whether learner-supplied contexts become clearer with time and in the knowledge that these will be used in a quiz by others, and whether definitions are copy-pasted from another source or truly data-constructed – as is encouraged by the imposition of a 100-character limit on both examples and definitions).

Of particular interest in the context of points raised above is the *Quiz avec de nouveaux contextes* ('Quiz with new contexts') provided as a further quiz option in the screen shown in Figure 9. A click of this button takes the learner out of Group Lex to a concordance exercise where these same words must be interpreted and inserted into novel contexts from the corpus of *Le Monde* newspaper writings already mentioned (Figure 10). Again, this 'multi-concordance' activity is a hands-on, semi-controlled, purposeful employment of the concordance information and format that learners and teachers alike have shown enthusiasm for—and research has confirmed the value of (Cobb 1999).

The CEGEP where Group Lex will be deployed with eight sections of lower intermediate French learners has asked for a full French translation of the interface, which is about 90% complete at time of writing, including concordancers, corpora, dictionaries, text-to-speech, and a cross-linguistic interface between the PHP (Figures 1 and 2) and PERL (Figure 3) programming environments. This interfacing is in keeping both with Quebec government stipulations and with a pedagogical desire to present students with an integral L2 environment. French Group Lex went into trial use in the winter session of 2013, and will be modified in line with learner observations and feedback. This, then, is another data-driven application that has been adapted to French fairly simply.

#	Qz	NOUVEAU MOT	EXEMPLE	CLASSE DE MOT	DÉFINITION	GROUPE	ÉTUDIANT	DATE/HEURE SOUMISE
1		**française**	Elle est **française** mais née au Canada	Adj	French (femininine)	Arts	tom	2012.12.21 18.57
2		**tôt**	Daisy s'était levée **tôt** ce matin de printemps.	Adv	Entre 4 heures et 7 heures du matin.	Arts	zz_tom	2012.12.14 10.50
3		**étudié**	Elle avait **étudié** le latin	Verb	studied	Arts	tom	2012.12.16 17.49
4		**tête**	J'ai mal à la **tête** quand je regarde la télévision longtemps.	Noun	Partie du corps qui contient le cerveau.	Interêt_général	rosecegep	2012.12.14 10.53
5		**tâche**	Les étudiants avaient à finir leurs **tâches** avant de quitter la salle de classe.	Nom	Leur travail	Interêt_général	tom	2012.12.21 19.08
6		**déjà**	J'ai **déjà** vu ce film.	Nom	Antérieurement	Interêt_général	tom	2012.12.21 22.58
7		**sécurité**	Elle se sentait en toute **sécurité** en ville la nuit.	Nom	Security, or better safety	Interêt_général	tom	2012.12.26 13.30
8		**sécurisé**	Les portes sont **sécurisées** avant le décollage	Nom	Fermées	Interêt_général	tom	2012.12.26 18.55
9		**lambeau**	Pendant plusieurs jours de suite des **lambeaux** d'armée en déroute avaient traversé la ville.	Noun	Morceau	Interêt_général	mb_6410	2009.12.31
10		**abeille**	Rien n'empêchait cette inlassable **abeille** travailleuse de raffiner son miel	Nom	Insecte qui fabrique du miel (a honey bee)	Interêt_général	tom	2013.01.03 13.02
11		**partions**	Il a fallu que nour **partions** de bonne heure	Adv	Quitter un endroit, s'en aller, provenant de partir	Interêt_général	zz_tom	2013.03.23 12.56

Figure 8. The main screen of a Group Lex

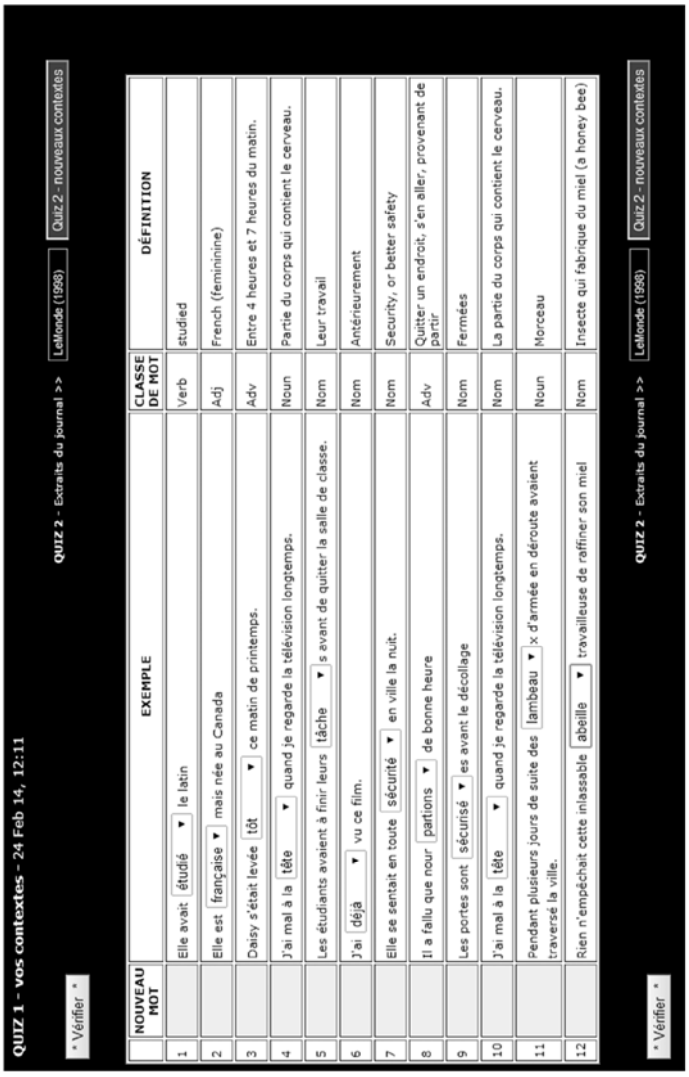


Figure 9. Group Lex Interactive Quiz

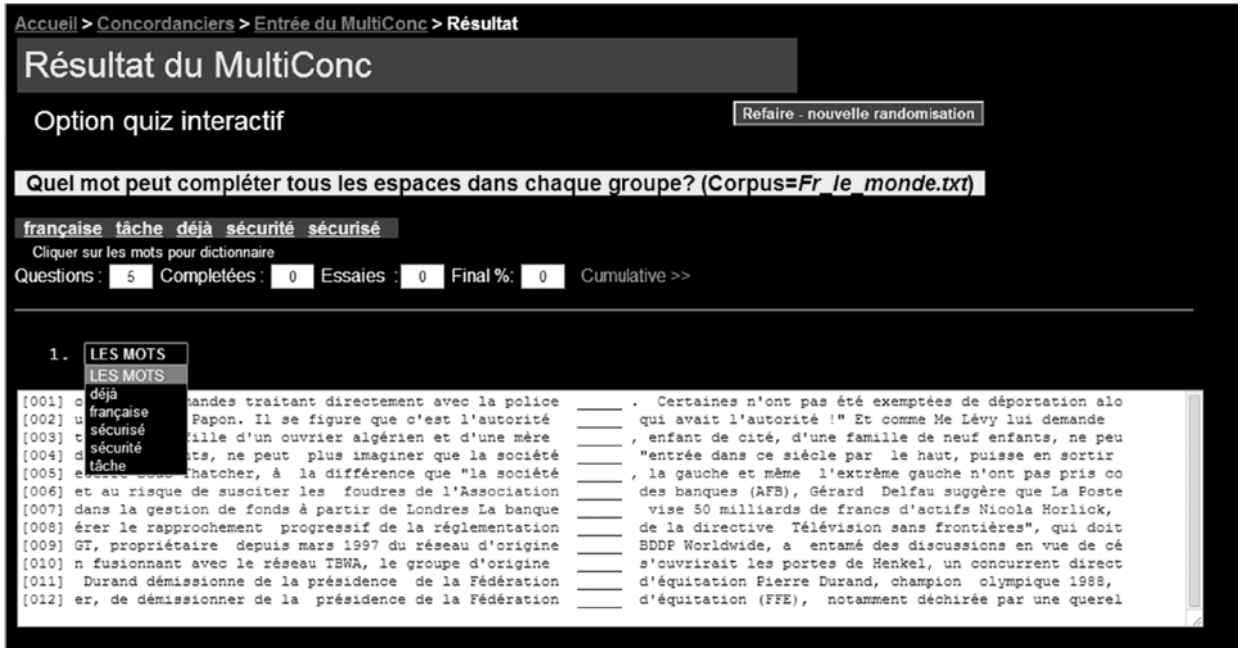


Figure 10. Multiconcordancing: transfer of word knowledge to novel contexts

So far we have seen that a substantial amount of DDL work can be set up for learners to do just using texts (Dictator, Range for Texts, and Group Lex) or small corpora (Hypertext, ConcordWriter, Concordance Feedback, and Range for Corpus). A great deal more than this is possible, however, with more and better organised versions of the first 'D' in DDL, data, the data on which learning will proceed. This can be shown by what has been done in English with its longer experience in this approach, or at least in making it available to language learners.

What is needed in French language DDL?

The French language adaptations described above have already hinted at some of the elements that are needed to make DDL as viable in French as it has been in English. These elements, it will be argued in the following section, include access to a broader range of corpora, access to larger corpora, pedagogical adaptation of large corpus data, particularly in the form of frequency lists, and pedagogical exploitation of these frequency lists.

A broader range of corpora

The simple concordance-based vocabulary review game shown in Figure 2 made the point that DDL activities can be adapted to suit a wide range of proficiency levels and learner interests including those in the early phases of language acquisition. We can now make the additional point that it is not possible to offer this activity to French learners, for the simple reason that no corpus of simplified French materials is available. French corpora on the whole are quite difficult to lay hands on. The *Le Monde* corpus on Lextutor, while adequate for certain purposes such as showing typical collocations on ConcordWriter or correcting typical errors on Concordance Feedback, is not a sampled corpus like the BROWN or any of the larger, more recent corpora. That is, no attempt has been made to represent a variety of written and spoken genres. Thus the range analysis shown above for *analy** showing the form's distribution across fiction, press, and academic sub-divisions in the one million-word BROWN corpus could not be performed using the *Le Monde* corpus despite its equal size, since the latter contains journalism only. This problem is not limited to French: there is a serious shortage of well sampled but modestly sized (for smooth running on the web) corpora in French, Spanish, and German. (Lextutor assistants have made up for this inadequacy by constructing BROWN-like

corpora for the latter two, the Braun in German and the Bruno in Spanish, playing mnemonically upon the original name, but have not yet done so for French).

However, it is not only small and web-runnable corpora that are needed to create a viable DDL in French, but also access to large corpora such as those that English practitioners have enjoyed since the creation in 1994 of the 100-million-word, 100 sub-divisioned, BRITISH NATIONAL CORPUS (1994). The BNC's sub-divisions are small corpora in their own right. English users of Lextutor's concordance programme can, for example, inspect a word or expression through a series of smallish corpora including the generalist BROWN, the legal or medical sub-divisions of the BNC, the simplified story collections, and others. But while Lextutor is not well adapted to run the BNC as a whole (for this see BNC-Web at http://bncweb.lancs.ac.uk), the broader BNC has played an enormous role in the development of DDL in English.

The BNC as a basis for frequency information

To be clear, there is no shortage of large and representative French corpora, such as the CORPUS DE LA PAROLE (http://corpusdelaparole.tge-adonis.fr/) and FRANTEXT (http://www.frantext.fr/), and thanks to a reviewer of this chapter for suggesting others including LEXIQUM (http://atour.iro.umontreal.ca/cgi-bin/lexiqum), CORPUSEYE (http://corp.hum.sdu.dk/cqp.fr.html), and PFC (http://www.projet-pfc.net/), among others. None of these, however, appear to make their corpora available to teachers or developers either for integration into the type of learner activities illustrated above (although PFC does offer corpus-based phonology activities for FLS learners) or for further pedagogical development. There is a serious shortage of public access to these corpora, and subsequently of opportunities for pedagogically oriented adaptation and formatting, as has been commented upon by many. Nicolas *et al*. (2002) lament the poor state of public access to French corpora; Véronis (2000: 2) discusses aspects of "*le retard du français*" in the formatting of its corpora for pedagogical and other practical purposes. What type of 'pedagogical development' or corpus information is needed? A model is Nation's (2006) work with the BNC frequency lists.

The frequency lists extracted from the BNC by Leech *et al*. (2001) had already been tagged for part of speech and lemmatised, so that rather than producing separate frequency counts for *move*, *moved*, and *moving*, for example, a combined count of all forms together was produced. To this, Nation (2006) added the more pedagogically pertinent grouping concept of

'word family'. This is an expansion of the lemma to include transparent derivations generally involving a change in part of speech, such as *movement* and *mover* in the case of *move*. This expansion was made for pedagogical reasons rather than purely linguistic ones, the idea beubg that transparent derivations should not normally create an additional learning burden for a learner who already knows *move* plus some basic morphologies (as defined by Bauer & Nation 1993, as part of the long lead-up to this work; see also http://www.lextutor.ca/morpho/mainfix). Nation and his programmers further devised a means to operationalise word families in the code of his version of the computer programme Range, such that any form of a word family will be picked up in a search if requested. This technique has been adapted throughout Lextutor routines. Its English Concordancer, for example, can assemble all the forms of any word, as shown in the output in Figure 11 for the *move* family. By contrast, the French concordances that Lextutor produces cannot output complete family groupings because the families have not been coded beyond the third thousand items. Note that in Figure 7, *prussien* and *prussiens* are presented as if they were completely unrelated words. In Figure 3, the information is assembled for *prestigieuses*, while any further information that might have been provided by *prestige*, *prestigieux* or *prestigieuse* is simply absent. Due to the lack of family-based or at least lemmatised lists, the whole analysis is limited to word forms. That said, families can be simulated to some degree using 'starts with' searches (Figures 4 and 5), such that 'starts with *prestig** will produce all singular, plural, masculine, and feminine forms. This, however, is a search that requires careful crafting from a sophisticated knowledge base that learners would not normally possess, and that could never be trusted to a pre-programmed mouse-click (cf. Figure 3). An incidental advantage of family searches is that they are economical in web programming terms, with more information generated per trip to the server.

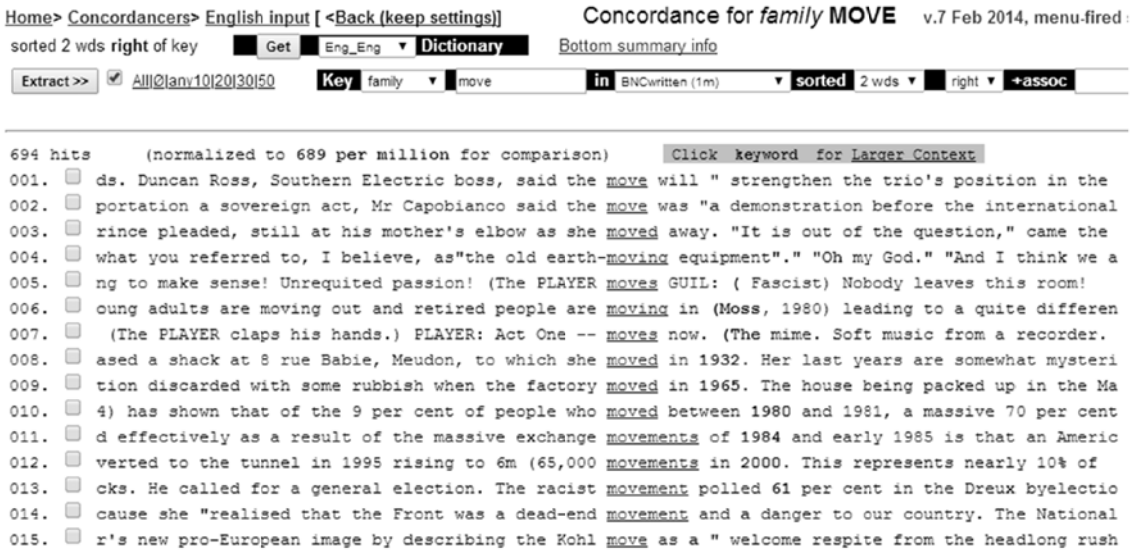


Figure 11: Keywords and collocates for whole families

Size matters

The main pedagogical benefit of a truly large corpus is that it can generate frequency lists that are reliable beyond just the highest frequency zones of a language. Almost any million-word collection of texts, whether of old letters or short stories, will produce similar frequency lists at the high frequency end of the language—the first two or three thousand word families—but, thereafter, size matters. The BROWN frequency lists were reliable up to about 3,000 word families, but after that the reduced number of occurrences of even medium-frequency items (discussed in Cobb 2007) made these lists unreliable, in the sense that different corpora generate different information. For example, the BROWN corpus has only 28 instances for all members of the sixth thousand-level word-family *stern*, while the equally sized BNC Written Sampler has 12; the BROWN has 21 instances for seventh thousand-level *peril*, the Sampler only five; and so on. Parity of frequency is only achieved with larger corpora.

Further, despite its name, a pedagogical frequency list is not only based on frequency, but also on range, that is on the number of sub-corpora the word family appears in. If, for example, all 21 instances of *peril* appear in adventure fiction, with no instances at all in newspapers, textbooks, biographies and other sub-corpora, then it is not necessarily an essential item for a learner to learn. The BROWN, as noted above, does have subdivisions but they are rather small (the entire corpus is just one million words). The BNC, by contrast, consists of 100 sub-corpora of a million words apiece, making it possible to build a frequency list based on reliable information about both frequency and range. In the lists compiled by Nation (see 2006), each item has its place based on the two considerations of range (or distribution) as well as raw frequency.

Having complete and reliable frequency information is extremely important for the development of DDL, mainly because it is needed to create the data that learners can use to engage in data-driven learning. Its importance can be seen clearly in the context of the Lextutor programme Vocabprofile. This is a frequency programme designed to analyse users' texts, but rather than giving the frequency of every item in the text itself, it gives the frequency classification by 1,000 word family groupings in the language at large, as determined by Nation's pedagogical adaptation of the BNC frequency list. Prior to the existence of the BNC, frequency lists were known to be reliable up to only about 3,000 words. This meant that even mid-frequency items, let alone low frequency or specialist items,

were simply designated unclassifiable or 'off-list.' This was something of a problem, pedagogically, since it is precisely the mid-frequency zone where vocabulary growth falls off for many learners. Laufer (2000) shows that vocabulary growth consistently drops at about 2,000 word families; Cobb (2007) proposes the mechanism behind this phenomenon; Schmitt and Schmitt (2012) argue for the unsuspected importance of the mid-frequency lexicon in L2 development.

With only rudimentary frequency lists, there was no way to identify sources of language that have lesser and greater amounts of mid-frequency items, to test learners and see how much of it they know, or to modify texts to provide more targeted sources for it. But when Nation's (2006) BNC-based lists were incorporated into Vocabprofiles, suddenly all this became possible. Nation's lists, familised and based on both frequency and range, were developed up to 14,000 word-family classifications, and subsequently expanded to 20,000. These extended lists were used to create complete and reliable tests of vocabulary size (Beglar & Nation 2007) and incorporated into Range and Vocabprofile analyses along with an Edit-to-a-Profile feature allowing teachers to scale linguistic data up or down according to their learners' levels. Figures 12a and 12b show the lexical profile of a lexically rich 826-word newspaper text from the Canadian political satirist Rex Murphy. The main part of the frequency analysis appears on the right side of the text with a different colour for each frequency increment, shown against a black background (just as it appears live in a Demo at http://www.lextutor.ca/vp/bnc). In the first row, we see that 226 of the families in the text are at the K1 level (i.e. among the 1,000 most frequent families) and that these families account for 80.89% of all words in the text. The off-list or unclassifiable component (using the 20 frequency-levels scheme) is just 1.2%, and includes interesting but unessential nonce items like 'documentarian' and 'pitchman.' The same text run through earlier, pre-BNC versions of the software produced 9.9% unclassified items (as can be seen by running the same text through the archived programme at http://www.lextutor.ca/vp/eng). The difference between the two, of 8.7% of the 826 words (or about 70 words), is an interesting slice of the elusive mid-frequency vocabulary, nicely arranged by family occurrences as shown in Figures 12a-b and giving something of the flavour of the items involved. The goal of identifying this lexical zone is of course to provide learners with the data that they need for further predictable growth at different stages of learning.

Freq. Level	Families	Types	Tokens	Coverage (tokens)%	Cum%
K1 Words :	226	274	673	80.89	80.89%
K2 Words :	42	46	53	6.37	87.26%
K3 Words :	18	20	28	3.37	90.63%
K4 Words :	14	14	14	1.68	92.31%
K5 Words :	5	5	5	0.60	92.91%
K6 Words :	9	9	11	1.32	94.23%
K7 Words :	4	4	4	0.48	94.71%
K8 Words :	5	5	5	0.60	95.31%
K9 Words :	4	4	4	0.48	95.79%
K10 Words :	5	5	5	0.60	96.39%
K11 Words :	5	5	5	0.60	96.99%
K12 Words :	1	1	1	0.12	97.11%
K13 Words :	3	3	3	0.36	97.47%
K14 Words :	3	3	3	0.36	97.83%
K15 Words :	5	5	5	0.60	98.43%
K16 Words :	2	2	2	0.24	98.67%
K17 Words :				0.00	98.67%
K18 Words :				0.00	98.67%
K19 Words :	1	1	1	0.12	98.79%
K20 Words :				0.00	98.79%
Off-List:	?	10	10	1.20	100.00%
Total	352+?	447	832	100%	100%

Well, it was a narrow escape. But we did it. Canadians have preserved their liberties and independence against the always rapacious American beast.

We knew there were powerful elements in the United States that wanted us to kowtow and genuflect to a simplistic worldview, that knuckle-dragging Good-versus-Evil script they have been remorselessly propagandizing all over the world since 9/11. They have been trying to drag Canada into this simpleton's game for years, mauling truth […]

See Appendix for full text.

Figure 12a. Coding text richness with Vocabprofile

BNC-4,000 types: [fams 14 : types 14 : tokens 14] bulbs_[1] concluded_[1] construction_[1] cowboy_[1] distract_[1] essence_[1] flattered_[1] knuckle_[1] liberties_[1] preserved_[1] presidential_[1] sermons_[1] surplus_[1] wit_[1]

BNC-5,000 types: [fams 5 : types 5 : tokens 5] beast_[1] ego_[1] imperialism_[1] millionaire_[1] vacant_[1]

BNC-6,000 types: [fams 9 : types 9 : tokens 11] al_[1] bland_[1] celebrity_[3] defied_[1] mighty_[1] notch_[1] patented_[1] simplistic_[1] sternly_[1]

BNC-7,000 types: [fams 4 : types 4 : tokens 4] battalions_[1] fahrenheit_[1] municipal_[1] perils_[1]

BNC-8,000 types: [fams 5 : types 5 : tokens 5] banishing_[1] formidable_[1] nuance_[1] parable_[1] template_[1]

BNC-9,000 types: [fams 4 : types 4 : tokens 4] conspiracy_[1] decoder_[1] entrepreneur_[1] remorselessly_[1]

BNC-10,000 types: [fams 5 : types 5 : tokens 5] assorted_[1] awe_[1] caricature_[1] conquest_[1] download_[1]

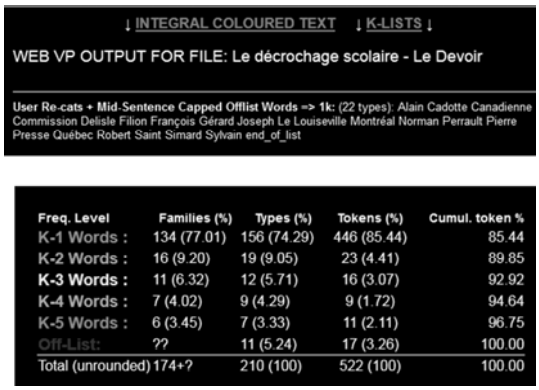
BNC-11,000 types: [fams 5 : types 5 : tokens 5] busybodyism_[1] insolence_[1] mauling_[1] posse_[1] subtlety_[1]

Figure 12b: Mid-frequency component (from a different part of the analysis shown in Figure 12a)

In terms of the organisational scheme depicted in Figure 1, this identification of suitable data for learning will most likely be done by teachers or other experts on behalf of learners, but learners can also be given a role. 'Learner Vocabprofiling' is an increasingly widespread use of this software, according to an informal survey of Lextutor users at the AAAL (American Association of Applied Linguistics) conference in 2011, normally with a view to encouraging awareness of the different lexical zones of English. Here is a more task-oriented approach that has been successful in English: learners determine their level using Beglar and Nation's (2007) BNC-based vocabulary size test, which provides scores in terms of the same 1,000-family increments as shown above. They then find texts for themselves and their classmates bearing suitable proportions of target 'next level up' lexis by searching on the World Wide Web. For example, learners who test at the fourth thousand-word level will look for texts in their areas of topic interest that carry 5% or more fifth to eighth thousand-level items, and so on. Or this could be done in a more formal manner through a database of texts especially prepared by lexical level, such as Carnegie-Mellon's REAP (Reader-Specific Lexical Practice for Improved Reading Comprehension; Eskenazi & Juffs 2012) database.

A rudimentary French version of the 'Vocabprofil' routine exists. Although able to generate significant distinctions between the lexis of beginner and intermediate learner language productions (Ovtcharov *et al*. 2006; Lindqvist 2010), it includes only three thousand-level categories and typically leaves 7% or more unclassified items (as shown in Figure 13, or run live at http://www.lextutor.ca/vp/fr). In other words, the frequency lists used in the French version of Vocabprofile, which were the best pedagogical lists available 10 years ago (Goodfellow *et al*. 2002; Jones 2002), basically divide the lexis of a French text into a high frequency zone on the one side and an undifferentiated zone of medium and low-frequency items on the other. While this was a promising beginning, frequency work in French (at least of the pedagogical variety) has now fallen quite far behind its English inspiration. Even so, few FSL teachers seem aware of any recent work on frequency in the language they teach; in a recent show of hands at a conference of FSL teachers in Canada, the term 'frequency list' equated to the now very dated *Français fondamental* (Gougenheim *et al*. 1964). The recently published *Frequency Dictionary of French: Core Vocabulary for Beginners* (Lonsdale & Le Bras 2009) may provide some motion forward on the frequency file, but since it classifies only 5,000 word families, is based on a corpus of only 23 million words, and is encoded in a CD-ROM which will make word-list extraction

difficult, it seems unlikely to do all that is needed to advance French-language DDL.



Des fonds sont débloqués pour réduire le décrochage scolaire au secondaire

Le taux de décrochage dans les écoles secondaire des milieux défavorisés devrait être réduit de 10%. Le gouvernement québécois investira 125 […]
See Appendix for full text

Figure 13. Vocabprofil for French newspaper text 'Taux de décrochage' (*La Presse*, Montreal, 2006)

However, the uses found in English applied linguistics for the large and accessible BNC does not end with better single-word frequency lists, as important as this is. There is a great deal more that a large corpus can deliver.

The BNC as a basis for phrase information

Assembling English corpora led to the discovery of the sheer amount of phrase repetition there is in English (Sinclair 1991). This is probably also the case in any language, although this is an empirical and possibly also a definitional question. For English, this realisation has brought about a number of important changes in how languages are taught, learned, and even conceptualised. At the most fundamental level, the Chomsky-inspired slot-and-filler model of language (if the slot in the sentence says 'noun,' then any noun will do) has now been largely replaced by an emphasis on whether the noun is idiomatic for the slot – in other words, whether the word is part of a recognisable phrase or multi-word unit. The acquisition of phrases in both L1 and L2, the reconceiving of grammar as the grammaticisation of phrases, and the need for learners to notice phrases in their input, have now 'revolutionised' ESL research and practice (Lewis 1993; Wray 2002). Very little of this has seemed important so far in

French applied linguistics. However, even in English it has until recently been a revolution without a syllabus.

The problem with identifiable phrases (as pulled out of a corpus by a computer programme like N-Gram, at www.lextutor.ca/tuples/) is the sheer number of them, and the relative infrequency of most of them. Both facts point to the sheer impossibility of teaching them all or drawing learners' attention to them all, or even letting them find them for themselves, DDL-fashion. Access to the BNC has to some extent reduced this problem. The size of the BNC, and its partitioning into range-checkable sub-corpora of suitable size, has finally allowed the search for a phrasal syllabus to begin. There have been two major BNC-based trial formulations of this search to date. Shin and Nation (2008) identified 891 high-frequency multi-word units which, if counted as single words, would belong in the most frequent four thousand words of English (84 in the first 1,000, 224 in the second, 259 in the third, and 324 in the fourth). In a larger scale analysis, Martinez and Schmitt (2012) ran the computer programme Wordsmith 5.0's n-gram extractor on the full BNC corpus in a run of just under four days and nights, creating a candidate list of multiword units that, when sampled and assessed by humans, led to the creation of a list of 505 items with frequencies that would suggest their insertion into thousand-lists one through five (32 in the first 1,000, 75 in the second, 127 in the third, 156 in the fourth, and 97 in the fifth). There are differences in the way the two studies defined the phrases to be included: Martinez and Schmitt looked for phrases with independent or non-compositional meaning, while Shin and Nation did not, so there are issues about what the phrasal syllabus will eventually look like.

But at present either Shin and Nation's or Martinez and Schmitt's version of this syllabus is interesting to learners and lends itself to hands-on DDL activities. One is to set learners the task of determining what proportion of all the instances of (for example) the word *course* throughout a medium-sized corpus is found in the multiword unit *of course* (645 concordance lines out of 757, in the case of the graded readers corpus on Lextutor).

The important point is that access to the BNC has made this large-scale work beyond the word form possible; and there does not appear to be corresponding work in French language pedagogy, almost certainly due to the lack of a large and accessible corpus. Lextutor does however enable more modest work with French phrases.

As they await large corpus breakthroughs, French teachers who are interested in working with multi-words in DDL will find text-based recurring-phrase activities easy to come up with. For example, learners

read a text such as the one shown in Figure 11 on school drop-out rates (*décrochage*), first for meaning, and then in pairs they underline all the phrases that recur at least twice and include two or more content words (not just *de la* or people's names). Then they confirm their work by pasting it into a programme that pulls out phrases automatically (such as Lextutor's Compleat Lister, at http://www.lextutor.ca/freq/compleat_lister) and compare their results to the computer's (see Figure 14).

4_millions_de_dollars	2_taux_de_décrochage
2_cinq_ans	2_mères_adolescentes
2_commissions_scolaires	2_rester_à_l`école
2_milieux_défavorisés	

Figure 14. The *décrochage* article's recurring strings

Teachers report that learners find it interesting the computer has noticed phrases that they had not. Indeed, the fact that a computer notices things about language that a human does not is a core DDL concept. This is thus an awareness-raising activity, which can be pursued further by having students recreate a summary of the text through a reconstruction from the phrases. The fact that this is possible shows them that the phrases are central to the text's meaning, and this ideally transfers to watching for recurring phrases in all their exposures to the L2.

Conclusion

The first precondition for a French DDL will be to develop or get access to more kinds and sizes of corpora. Access to a corpus like the BNC will make it possible to perform some version of the BNC-based work that has been done in English on word and phrase frequency, incorporating considerations of range. One of the reviewers for this chapter included in his or her comments an interesting account of the reasons there is no national corpus of French comparable to the BNC (potential disagreements about size, composition, tagging, and many other components of the task). And yet a large corpus-based syllabus of high and medium frequency words and phrases is vital if French as a second language is to develop a full DDL component. Following that, the next priority is probably to obtain or develop corpora at different levels of language proficiency, along the lines of the graded corpus on Lextutor. The vast majority of language learners are low-intermediate (who know about 1,000 word families, and can make their way through a newspaper story with lots of dictionary use). DDL is only interesting if it can be used with a broad range of learners,

and the low end in French will need work. A final priority is to have a number of smallish written corpora (1 to 3 million words) that can be used for fast processing online in programmes like ConcordWriter or Hypertext.

The second precondition will be to develop tools that French learners can use for making sense of corpus data that match their levels, tasks and interests. A sample of the tools developed on Lextutor for English were presented above, each adapted to a learner need and each with a strong track record with learners. Many of these tools are hospitable to French as a second language, and FSL practitioners are welcome to use them or add to them. But there are probably other approaches that will be better suited to a different language type and learner type. A way to proceed is to check what French learners are doing in French on their smart phones and see if there are any ideas there for learning tools. The point is to build on what they are already interested in, and right now 'data' is a very positive word for learners.

Bibliography

Bauer, L. & P. Nation 1993, "Word families", *International Journal of Lexicography* 6(4), 253-279

Beglar, D. & P. Nation 2007, "A vocabulary size test", *The Language Teacher* 31(7), 9-13

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finnegan 1999, *Longman Grammar of Spoken and Written English*, Harlow: Pearson Education

Biggs, D. & M. Dalwood 1976, *Les Orléanais ont la parole*, London: Longman

Boulton, A. 2010, "Data-driven learning: taking the computer out of the equation", *Language Learning* 60(3), 534-572

Boulton, A. 2011, "Data-driven learning: the perpetual enigma", In S. Goźdź-Roszkowski (ed.), *Explorations across Languages and Corpora*, Frankfurt: Peter Lang, 563-580

British National Corpus, version 1 1994, distributed by Oxford University Computing Services on behalf of the BNC Consortium, http://www.natcorp.ox.ac.uk/

British National Corpus, BNC Sampler XML version 2005, distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/

Cobb, T. 1997, "Is there any measurable learning from hands-on concordancing?" *System* 25(3), 301-315

Cobb, T. 1999, "Applying constructivism: a test for the learner as scientist", *Educational Technology Research and Development* 47(3), 15-31

Cobb, T. 2007, "Computing the vocabulary demands of L2 reading", *Language Learning and Technology* 11(3), 38-63

Cobb, T. 2009, "Internet and literacy in the developing world: delivering the teacher with the text", In K. Parry (ed.), *Literacy for All in Africa. Beyond the School* (vol. 2), Kampala: Fountain/African Book Collective, 627-645

Ellis, N. & R. Schmidt 1997, "Morphology and longer distance dependencies: laboratory research illuminating the A in SLA", *Studies in Second Language Acquisition* 19, 145-171

Eskenazi, M. & A. Juffs 2012, "Information retrieval for reading tutors", *Encyclopedia of Applied Linguistics*, New York: Wiley, doi: 10.1002/9781405198431.wbeal0536

Gaskell, D. & T. Cobb 2004, "Can learners use concordance feedback for writing errors?", *System* 32(3), 301-319

Goodfellow, R., M.-N. Lamy & G. Jones 2002, "Assessing learners' writing using lexical frequency", *ReCALL* 14(1), 133-145

Gougenheim, G., R. Michea, P. Rivenc & A. Sauvageot 1964, *L'élaboration du français fondamental: étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris: Didier

Horst, M., T. Cobb & I. Nicolae 2005, "Expanding academic vocabulary with an interactive on-line database", *Language Learning & Technology* 9(2), 90-110

Hulstijn, J. & B. Laufer 2001, "Some empirical evidence for the involvement load hypothesis in vocabulary acquisition", *Language Learning* 51(3), 539-558

Johns, T. 1986, "Micro-Concord: a language learner's research tool", *System* 14(2), 151-162

Johns, T. 1997, "Contexts: the background, development and trialling of a concordance-based CALL program", In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds.), *Teaching and Language Corpora*, London: Longman, 100-115

Jones, G. 2002, *Compiling French Word Frequency Lists for the Vocabulary Assessment Tool*, Open University technical document, http://www.lextutor.ca/vp/fr/glynn_jones.html

Laufer, B. 2000, "Task effect on instructed vocabulary learning: the hypothesis of 'involvement'", *Selected Papers from AILA '99*, Tokyo: Waseda University Press, 47-62.

Leech, G., P. Rayson & W. Wilson 2001, *Word Frequencies in Written and Spoken English: Based on the British National Corpus*, London: Longman

Lewis, M. 1993, *The Lexical Approach*, Hove: Language Teaching Publications

Lindqvist, C. 2010, "La richesse lexicale dans la production orale de l'apprenant avancé de français", *La revue canadienne des langues vivantes* 66(3), 393-420

Longman 2011, *Longman Dictionary of Contemporary English*, Harlow: Pearson Education

Lonsdale, D. & Y. Le Bras 2009, *Frequency Dictionary of French: Core Vocabulary for Beginners*, New York: Routledge

Martinez, R. & N. Schmitt 2012, "A phrasal expressions list", *Applied Linguistics* 33(3), 299-320

McCarthy, M., J. McCarten & H. Sandiford 2006, *Touchstone*, New York: Cambridge University Press

Nation, P. 2006, "How large a vocabulary is needed for reading and listening?", *Canadian Modern Language Review* 63(1), 59-82

Nation, P. & A. Heatley 2002, *Range: A Program for the Analysis of Vocabulary in Texts*, http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Nesselhauf, N. 2005, *Collocations in a Learner Corpus*, Amsterdam: John Benjamins

Nicolas, P., S. Letellier-Zarshenas, I. Schadle, J. Antoine & J. Caelen 2002, "Towards a large corpus of spoken dialogue in French that will be freely available: the 'Parole publique' project", In *Proceedings of the 3rd International Conference on Language Resources & Evaluation, LREC'2002*, Las Palmas de Gran Canaria, 649-655

Ovtcharov, V., T. Cobb & R. Halter 2006, "La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière", *Revue canadienne des langues vivantes* 63(1), 107-125

Pinker, S. 1994, *The Language Instinct: How the Mind Creates Language*, New York: William Morrow

Schmitt, N. & D. Schmitt 2012, "A reassessment of frequency and vocabulary size in L2 vocabulary teaching", *Language Teaching*, doi: 10.1017/S0261444812000018

Shin, D. & P. Nation 2008, "Beyond single words: the most frequent collocations in spoken English", *ELT Journal* 62(4), 339-348

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

Véronis, J. 2000, "Annotation automatique de corpus", In J.-M. Pierrel (ed.), *Ingénierie des langues*, Paris: Hermès, 111-129

Wray, A. 2002, *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press

Appendix

Full Rex Murphy text profiled in Figure 9

Well, it was a narrow escape. But we did it. Canadians have preserved their liberties and independence against the always rapacious American beast.

We knew there were powerful elements in the United States that wanted us to kowtow and genuflect to a simplistic worldview, that knuckle-dragging Good-versus-Evil script they have been remorselessly propagandizing all over the world since 9/11. They have been trying to drag Canada into this simpleton's game for years, mauling truth and banishing nuance with a continuous stream of invective posing as reason, and caricature passing itself off as accuracy.

It's a difficult thing to resist the mighty United States at any time, and especially difficult in all the dust and storm of a national election. But we did it.

It was a close-run thing. But on Monday night, Canada fought back and won. On Jan. 20, just three days before our vote,

Michael Moore, entrepreneur, fabulist, philosophe, issued a broadside to the citizens of this country warning us sternly, and with the imperious irony of which he is so fully a master, against the perils of electing a Stephen Harper government: Do you want to help George Bush by turning Canada into his latest conquest? Is that how you want millions of us down here to see you from now on? The next notch on the cowboy belt? I was worried at first that the subtlety of the pitch might obscure its wonderful impertinence — worried that the charm of Mr. Moore's address might distract Canadians from the consideration that an American millionaire celebrity pitchman was interfering in, and attempting to influence, the Canadian vote.

I was worried, too, that this one-man shock-and-awe "documentarian" might be leading a charge, that the other bright bulbs of international busybodyism were massed behind his formidable massed behind. Was Sean Penn on the way to monitor the vote in Etobicoke? Was he planning one of his patented fact-finding junkets like the visits that brought such comfort and peace to the citizens of Baghdad? I could see the headlines: Penn in Halifax. Visits Bar. Reads Construction-Site Posters. Warns Harper is Christian. Says "God Bless Canada."

Well, that didn't happen. We're were spared the fast-food internationalism of Mr. Penn, and that probably meant we were spared assorted sermons from Alex Baldwin, Janeane Garofalo, Al Franken and that whole posse of celebrity dilettantes who see the whole world as an audience for their inch-deep, paint-by-numbers, cause-a-day homilies.

Maybe they were off somewhere saving a seal.

Or, what is much more likely, maybe he concluded there was really no need for the secondary battalions. We, the respectful, bland and polite citizens of a country that is really only a farm team for the U.S. entertainment industry — hello Céline, Jim, Dan and Avril — would naturally be flattered into sheer insensibility that the portentous Mr. Moore even knew we were having an election. He has a taste for insolence, referring to Stephen Harper, who has more brain than Michael Moore has girth, as someone "who should be running for governor of Utah," and whose election would "reduce Canada to a cheap download of Bush & Co."

One size fits all — that's our Mikey. Because he thinks he has a problem with George Bush, that must be the script for the rest of the world. This is the very essence of imperialism. To believe that your story is everyone else's. To believe that your political drama is the template for every other political drama in the whole wide world. Michael Moore could go to Fogo Island, Nfld., for the municipal elections and find them a perfect parable of the Halliburton super-conspiracy. He'd see Dick Cheney's influence in the selection of the town clerk.

Full French text profiled in Figure 10

Des fonds sont débloqués pour réduire le décrochage scolaire au secondaire.
Le taux de décrochage dans les écoles secondaire des milieux défavorisés devrait être réduit de 10%. Le gouvernement québécois investira 125 millions de dollars en cinq ans afin d'atteindre cet objectif. «Avec 36,6% de décrochage dans les

milieux défavorisés, plus d'un jeune sur trois quitte l'école sans un diplôme qui pourrait le faire sortir de la pauvreté et de l'exclusion», a résumé lundi le ministre de l'Éducation, Sylvain Simard, en annonçant le programme de lutte au décrochage. En fait, le ministre a ciblé 199 écoles secondaires du Québec où le taux de décrochage est largement supérieur à la moyenne. Il s'agit d'écoles situées pour la plupart dans des milieux défavorisés. Ces écoles sont réparties dans 54 commissions scolaires, 46 commissions de langue française et 8 commissions scolaires anglophones. Ces 199 écoles regroupent au total 36 124 élèves. L'argent additionnel qui sera versé à ces écoles, soit 25 millions de dollars par année pendant cinq ans, servira à prendre des mesures susceptibles d'aider les jeunes à rester à l'école et à compléter leur cours. Les fonds pourront servir à embaucher des aides pédagogiques (psychologues, pédagogues), à mettre à contribution les parents ou à fournir de l'aide additionnelle à certaines catégories d'élèves, par exemple, aux mères adolescentes. Une expérience a été menée depuis un an dans six écoles particulièrement défavorisées. À l'École Gérard-Filion, au centre-ville de Montréal, on a embauché des spécialistes en appui aux enseignants. Par contre, à Louiseville, en milieu rural, on a instauré un service supplémentaire d'autobus scolaires à 17h30, le soir, permettant aux élèves de retourner plus tardivement à la maison afin de compléter leurs devoirs et leurs travaux scolaires sous supervision des enseignants. «Le Québec a besoin de tous ses jeunes. Il faut donner aux jeunes des raisons de rester à l'école.

C'est nous qui avons le fardeau de la preuve; nous devons rendre l'école passionnante et stimulante», a dit le ministre Simard. Déjà, afin d'aider les écoles en milieux défavorisés, le gouvernement avait débloqué l'an dernier 50 millions de dollars supplémentaires, dont 10 millions de dollars pour celles situées à Montréal, 10,8 millions de dollars pour une aide alimentaire, et 28,7 millions de dollars pour le développement des maternelles et des services de gardes pour enfants de quatre ans. Les fonds de lutte contre le décrochage ont soulevé l'espoir dans les milieux concernés. Le porte-parole de la Commission scolaire de Montréal, Robert Cadotte, s'est réjoui que la somme mise à la disposition des écoles ne soit pas saupoudrée, mais investie dans les écoles qui présentent les problèmes les plus aigus. À Québec, le directeur de l'école Joseph-François-Perrault, Alain Saint-Pierre, a souligné que les fonds serviront à améliorer les services de son école qui reçoit des enfants provenant de familles à faibles revenus, de familles d'immigrants et qui accueille également des mères adolescentes.