Corpus for courses: Data-driven course design

Author
Thomas Cobb
Université du Québec à Montréal (UQAM)
Montréal,Canada
cobb.tom@uqam.ca
www.lextutor.ca

Data-driven learning (DDL) as a set of principles and technologies has a well-established role in language learning, and this paper shows how these can also be applied to language course design. If course materials are reformulated as a corpus, a number of ways become possible to bring learning research directly into the classroom. This article begins with a definition of terms, a review of the place of data-driven learning in language acquisition, and shows ways of applying DDL to the evaluation, design, and testing of language instruction. The context is the re-design of an ongoing adult second language reading course based on a collection of authentic materials found on the Internet. Principles and technologies of data-driven course design were used, first, to expose the weakness in this collection of materials as a course, and, second, to show concrete ways it could be substantially improved. All software involved in this work is publicly available.

L'apprentissage sur corpus, en tant qu'ensemble de principes et de technologies, joue un rôle bien établi dans l'apprentissage des langues, et cet article montre comment ceux-ci peuvent s'appliquer également à la conception de cours ou de cirruculum de langue. Si les supports de cours sont reformulés sous forme de corpus, un certain nombre de moyens deviennent possibles pour amener la recherche sur l'apprentissage directement dans la salle de classe. Cet article commence par une définition des termes, un examen de la place de l'apprentissage sur corpus dans l'acquisition de la langue, et montre des moyens d'appliquer cette approche à la conception, au testing, et à l'évaluation de l'enseignement des langues. Le contexte est la refonte d'un cours de lecture en cours pour adultes en langue seconde basé sur une collection de documents authentiques trouvés sur Internet. Les principes et technologies de la conception de cours sur corpus ont été utilisés, premièrement, pour exposer les faiblesses de cette collection de matériels en tant que cours, et, deuxièmement, pour montrer des moyens concrets de les améliorer. Tous les logiciels impliqués dans ce travail sont accessibles au public.

Mots-clés:
apprentissage basé sur les consultations d'un corpus, conception pédagogique, analyse de besoins, vocabulaire

Keywords:
corpus-based learning, data-driven learning, instructional design, needs analysis, text analysis, vocabulary

1. Background & proposition

Data-driven learning (DDL) is an input and comprehension-based approach to language learning, but with the proviso that second language (L2) input can be made more comprehensible to learners with its patterns exposed or highlighted by computer programs. An example of a DDL research finding is that new word meanings are inferred more successfully by learners consulting several examples of words assembled by software than by learners consulting single instances over a natural timespan (Cobb 1999). Other types of computer-assisted pattern exposure that are relevant to learning include text-to-speech algorithms that transform written language to spoken; search software that reveals distant and low-frequency collocations in a text (Greaves 2009); discourse tracking software that highlights the coherence threads through a text (Crossley et al 2016); error-tracking software for tagging corpora of learner writing (Granger 2003); corpus frequency lists for vocabulary sequencing and testing (Webb & Nation 2017); and many others. The power of data-driven learning in an array of applications involving corpus consultation by learners is shown in a meta-analysis of empirical studies performed by Boulton & Cobb (2017). The overall finding of a strong learning effect compared to other approaches is currently being unpacked in a series of finer cut studies (e.g. Lee et al. 2020).

The topic of this chapter, however, is not the learning power of DDL, but rather its potential use in course or curriculum design (hence data-driven course design, DDCD). The idea is that just as learners can profit from the computer’s ability to expose patterns in language, so at another level can course designers and teachers.

The first step in DDCD is to assemble learners’ inputs or materials into a corpus. A corpus is a large structured text with discernible sub-sections, which is representative of language beyond itself, whether on the scale of an entire language (the British National Corpus, BNC 2001; the Corpus of Contemporary American for U.S. English, COCA, Davis, 2008) or the language used by a defined subset of users. From a modest corpus of classroom materials, combined with insights from research on L2 acquisition, it will be possible to predict whether these materials are useful for these particular learners; what can be done to make them more so; what learning challenges and outcomes can be expected; what supplementary materials will be required; and what if anything will be examinable from what the learners have been exposed to.

The present case is an elaborated example focusing on an existing setting, namely an upper-intermediate remedial ESL reading course for adult Francophone learners in Quebec, Canada. The students were returning to the classroom to obtain high school graduation, which includes basic literacy in English. Their reading course consists almost exclusively of materials found on the internet, which is an increasingly common format for such courses. The typical design of such a course is to read, discuss, and answer questions about a series of such texts, then take a test based on a similar text. The present study stems from a request from the course designers for research-informed assistance to make the course more interesting and its outcomes more reliable. With data-driven tools, it should be possible, for example, to develop more varied content from the texts in the learning phase, and to check the similarity between course and test in the testing phase, and perhaps more.

To limit the topic, the focus of the intervention was primarily on the vocabulary component of the course, and the computer tools are an assortment from the author's website, Lexical Tutor (Lextutor, www.lextutor.ca), a suite of free online text analysis tools mainly relevant to learning English but with adaptation to French and other languages. Lextutor is a reverse engineering of some standard text analysis tools such that they can be (1) easily accessed by practitioners, (2) worked into coherent sequences of application, and (3) used to identify and resolve specific practical problems in language learning.

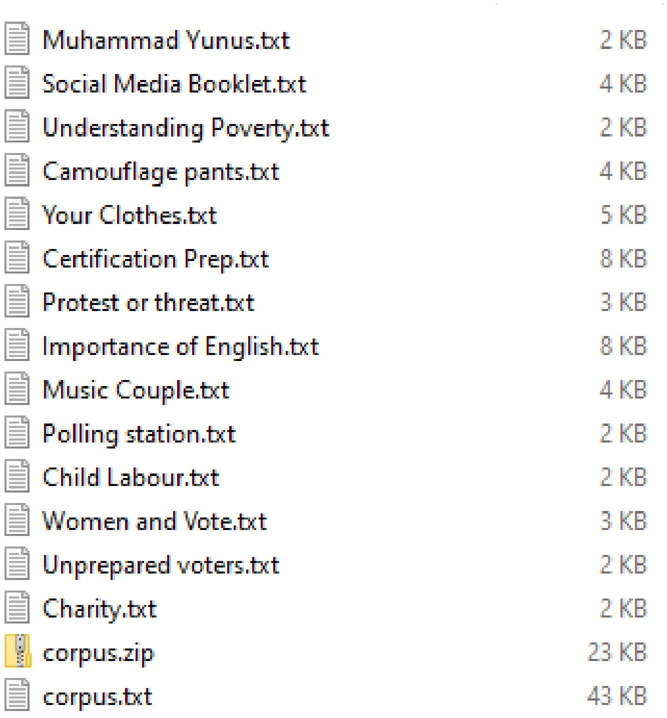
The plan of the paper is to show readers how to do the following: make and format a corpus; use vocabulary profiling tools in conjunction with vocabulary testing to judge the match between corpus and learners; and deploy a variety of related tools to improve the match between course, learners, and assessment. A sub-goal of the paper is to model one possible way of sequencing text analysis tools coherently. Until now, these have been validated separately and out of context, and, if ever used by practitioners, used separately. Though this is not an empirical study, the paper's hypothesis is that there is a *prima facie* case for designing reading courses as corpora, and the conclusion will include proposals for empirical validation and a discussion of the issues involved.

2. Making & analyzing a corpus

Putting together an Internet selection of reading materials has become easy to do and is an increasingly common practice. The advantages of doing this are obvious: A wide range of text types is readily available, in a size that can be read or re-read in one sitting, of a variety that is unlikely to provoke boredom, and that can be fairly easily matched to learners' environments and interests. Only a small amount of editing is likely to be required, e.g., to remove advertisements from news stories. But putting these texts together as a corpus can reveal some issues that are not obvious.

To become a smoothly functioning corpus, these texts have to be changed into text files (with *.txt* extension) so they do not carry formatting information. Doing this involves either saving the file as text or copy-pasting it into a text document. For the course in question, Figure 1 shows the text files involved, a de-formatted collection of news stories from the learners' environment. These are then combined in a compressed *zip* file, which maintains individual file identity, using either Mac or PC system software. They are further collected together as a large single text file by a Lextutor routine called *Corpus Builder* (www.lextutor.ca/cgi-bin/tools/corp_build). The corpus thus exists in three formats, as a collection of separate files and in two single file versions, and each is used for different purposes in the procedures that follow. The combined versions appear at the bottom of Figure 1. The combined corpus consists of 7,198 individual words in 1,460 word families distributed across fifteen 1000-family levels (in the word-frequency scheme of Nation 2016, to be discussed below). With data thus assembled, we are in a position to ask some interesting questions of it.

Fig. 1. Reading course as corpus



Muhammad Yunus.txt	2 KB
Social Media Booklet.txt	4 KB
Understanding Poverty.txt	2 KB
Camouflage pants.txt	4 KB
Your Clothes.txt	5 KB
Certification Prep.txt	8 KB
Protest or threat.txt	3 KB
Importance of English.txt	8 KB
Music Couple.txt	4 KB
Polling station.txt	2 KB
Child Labour.txt	2 KB
Women and Vote.txt	3 KB
Unprepared voters.txt	2 KB
Charity.txt	2 KB
corpus.zip	23 KB
corpus.txt	43 KB

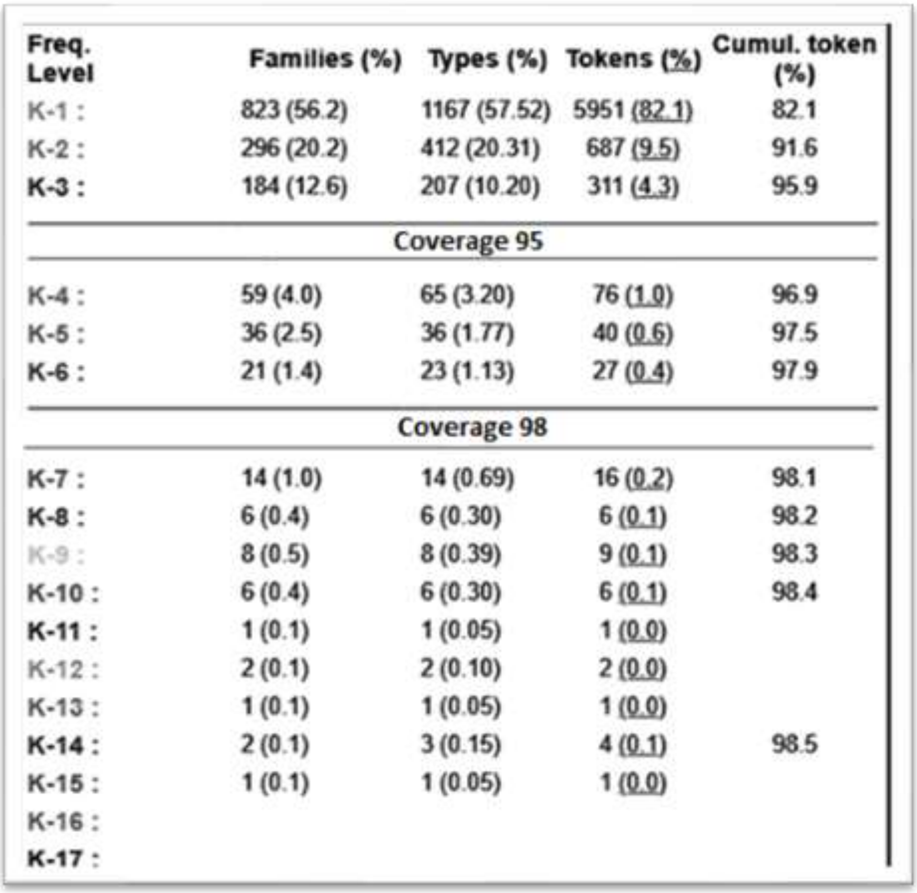
2.1 Can these learners read these materials?

The first thing to know about this corpus is whether it is readable by our learners, and following that whether it has anything for them to learn.

The primary assumption of the present analysis is that vocabulary knowledge plays a key role in reading comprehension and development (though grammar will also get some attention). A vocabulary focus could be seen as merely reflecting what text analysis happens to be good at, i.e., counting up the bits of text between spaces, which is true, but it is more than that. Of all the components of L2 reading comprehension (like topic familiarity, grammar knowledge, first language distance, first language reading ability, working memory capacity, and others), and despite the overlap between vocabulary knowledge and many of these, vocabulary knowledge reliably predicts the major share of the explained variance in reading comprehension (Bernhardt 2005). Thus readability is largely a matter of knowing vocabulary, with more common words normally being more readable than less common words, owing to the likelihood they have been met and processed more often and in a wider variety of contexts.

A tool for evaluating the vocabulary level of texts and small corpora is *Vocabprofile* (*VP*; at www.lextutor.ca/vp/comp). The single-file version of the corpus is used for this step in the analysis. VP traces every word in the corpus to its frequency rating in a much larger corpus of English (in this case, the BNC-COCA lists developed by Nation 2016). It then puts these ratings together into bands of 1,000-word families, as shown in Figure 1, where for example in our corpus, first-thousand families account for 82.1% of the total number of individual words, or word tokens. A word family is a headword (*read*) plus all its inflections (*reads, reading*) and its most obvious derivations (*reader),* as elaborated in Laufer & Cobb (2019). For brevity, the notation K-1 refers to words in the first 1,000 families (function words, proper nouns, and common lexical words); K-2 to the second (slightly less common words); and so on. The K-levels are arbitrary cut-offs (though they correspond roughly to what an avid learner might acquire in a year, and K-1 is a functional definition of the lexis of everyday speech).

Fig. 2. Lexical profile of the course corpus



Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token (%)
K-1 :	823 (56.2)	1167 (57.52)	5951 (82.1)	82.1
K-2 :	296 (20.2)	412 (20.31)	687 (9.5)	91.6
K-3 :	184 (12.6)	207 (10.20)	311 (4.3)	95.9
Coverage 95				
K-4 :	59 (4.0)	65 (3.20)	76 (1.0)	96.9
K-5 :	36 (2.5)	36 (1.77)	40 (0.6)	97.5
K-6 :	21 (1.4)	23 (1.13)	27 (0.4)	97.9
Coverage 98				
K-7 :	14 (1.0)	14 (0.69)	16 (0.2)	98.1
K-8 :	6 (0.4)	6 (0.30)	6 (0.1)	98.2
K-9 :	8 (0.5)	8 (0.39)	9 (0.1)	98.3
K-10 :	6 (0.4)	6 (0.30)	6 (0.1)	98.4
K-11 :	1 (0.1)	1 (0.05)	1 (0.0)	
K-12 :	2 (0.1)	2 (0.10)	2 (0.0)	
K-13 :	1 (0.1)	1 (0.05)	1 (0.0)	
K-14 :	2 (0.1)	3 (0.15)	4 (0.1)	98.5
K-15 :	1 (0.1)	1 (0.05)	1 (0.0)	
K-16 :				
K-17 :				

VP further matches K-levels to the percentages of words known that empirical research has shown to correspond to different comprehension levels. In Figure 2, ‘Coverage 95’ is drawn at the point where 95% coverage is reached, K-3 in this case. Empirical research has shown that 95% known-word coverage typically corresponds to ability to read and comprehend with the help of resources (dictionary, Google, etc.). Reading with 95% coverage also enables learners to acquire some of the remaining unknown vocabulary through contextual inference. ‘Coverage 98’ is reached at K-6 in the corpus, with 98% known words typically corresponding to independent reading and native speaker inferencing ability without resources (Laufer & Ravenhorst 2010; Schmitt et al. 2011). Extending the framework downward, reading with much less than 90% coverage is likely to be arduous, error prone, and lead to little further vocabulary growth through inference (Hu & Nation 2000). This leaves reading with 85%-95% known-word coverage as the zone where reading development is feasible and instruction justified. Without this minimal knowledge of what the words in a text mean, the usual ‘high level’ activities of the reading classroom (find the main idea, track the references and transitions, distinguish fact from opinion, articulate the implicit) are not doable (Alderson 1984).

VP analysis is thus interesting, but it would remain somewhat theoretical if our learners’ knowledge could not be brought into relation with the corpus profile. In practice, this can be done because the K-levels scheme has been employed in a number of vocabulary tests that fully correspond to the text profile framework. Several such tests, in a variety of formats produced by a variety of researchers, are available in multi-platform versions with score recording at www.lextutor.ca/tests. A pilot sample of 11 of the learners for whom these reading materials had been developed were tested with the most comprehensive of these tests, the Vocabulary Size Test (VST) developed by Nation & Beglar (2007). The VST tests meaning recognition by sampling 10 random words at each K-level, from K-1 (*jump, shoe, stone*) to K-14 (*plankton, skylark, beagle*). The score is a percentage of correct answers at each level. The scores in this case showed that six of the learners were strong all the way from K-1 to K-5 (*devious, threshold, veer*), meaning they could almost certainly read these texts easily (with almost 98% of words known), but with little to gain other than fluency practice, deepening of knowledge for words already known, and possibly a few new acquisitions from the 73 available beyond K-5 (all worthwhile achievements, but beyond the stated brief of this basic course – these learners should have had a different course).

Another two of the learners had strong knowledge at K-1, but less than 50% knowledge across the next three levels, meaning they would have knowledge of about 85% of the words in most of the corpus (K-1 is 82%, Fig. 2). These learners would probably have difficulty reading these texts by themselves, but could benefit from going through them with a teacher; they were in the instructable zone. For learners in the 85-95% known-word zone, these texts offer two important things: adequate or near-adequate contextual support for inferring the meanings of unknown words and plenty of unknown words to learn. The remaining four learners had less than 50% knowledge across the board, from K-1 to K-4, meaning they were in no position to profit from engaging with these materials. They should have been reading the most basic simplified readers.

To sum up, this course was effectively useless for all but two of the 11 class members that were tested. Would any of this discrepancy be noticed in a classroom without text profiling and vocabulary testing?

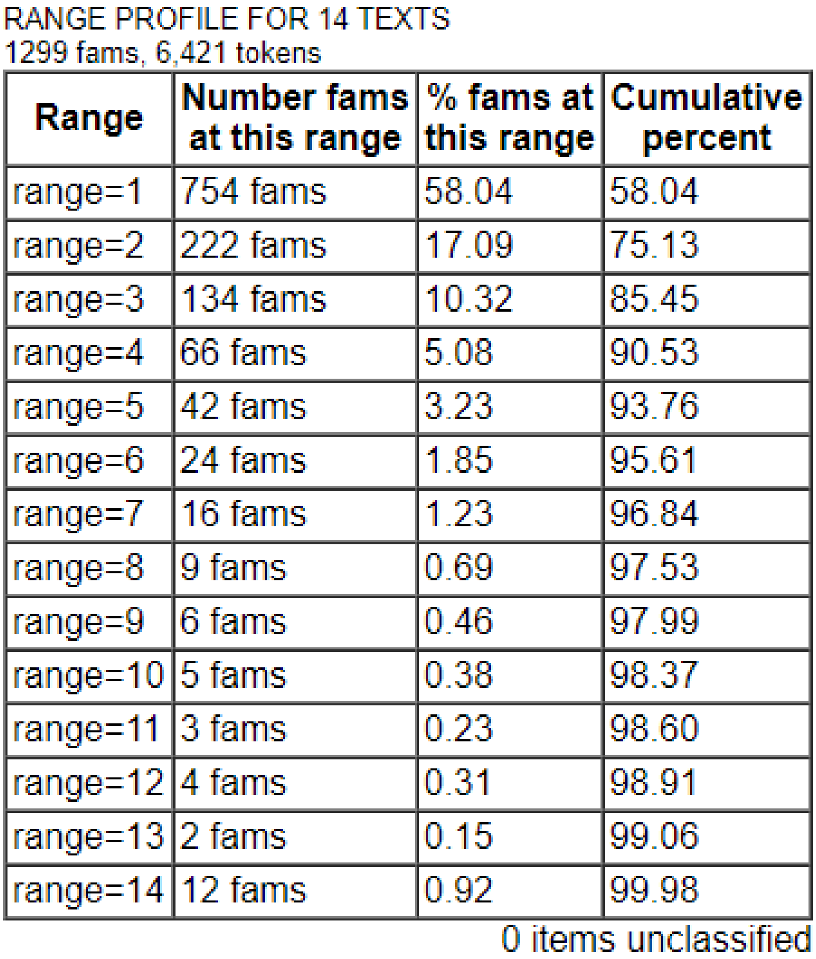
2.2 Can these learners learn anything from these materials?

From the profiling outcome, it is clear that little word learning or other skill development will result for most of these learners from engaging with these materials. For display purposes, however, let us suppose that the course corpus had been more or less within the learners' range – its vocabulary a bit more basic, or the learners a bit more advanced, so that lexical growth was a possibility for more of them. Even so, words are not learned unless encountered more than once or twice. Is that the case in this corpus?

To answer this question, we turn to another Lextutor routine called *Range for Texts* (www.lextutor.ca/cgi-bin/range/texts) which looks at word distributions over sets of texts. The zip file version of the corpus is entered into the program and then dissected family by family to see how many of the 14 texts each family appears in. (*Range* will count different family members as recurrences of the same word.) How often will the words in these materials be re-encountered? Empirical research suggests that 10 occurrences are typically needed for even rudimentary learning (Cobb 2007) and many more for productive use.

Figure 3 shows that for the 1,299 word families present in the course corpus, more than 75% have a range of just one or two out of 14, that is, they appear in in only one or two of the texts. Abundant research shows that with such infrequent occurrences they are unlikely to be noticed or learned. Of course, as a reviewer usefully points out, a new word may well appear multiple times in one or two texts, and even if not reinforced in a subsequent text within the course nevertheless establish an initial representation in the lexicon that could be developed at some future point. However, *Range* also tallies frequency data, though not text by text, and the average frequency of the words appearing in these 1 to 2 texts is just 1.8 occurrences (SD=2.15). In other words, the vast majority of word families appear just once per text. At the other end of the distribution, the roughly 18 words with a range of 12 to 14 are function words (*the, of, and*) without lexical meaning. This leaves a potentially learnable vocabulary load (content words encountered in from three to 11 of the texts) of just 305 word families. This contrasts to the 3,000 families needed to read these texts (at 95% coverage, with resources) and the learners' with knowledge of fewer than 1,000.

Such a distribution of learning opportunities is not sufficient for either the weak or the strong learners in the present sample. Strong learners need a rich reading diet to meet the words they do not already know; the VP analysis showed these learners meeting only 2.5% unknown items from 1,299 families, or 33 families. Weak or beginning learners need to meet high frequency words over and over to start building a lexicon, which *Range* analysis shows can not happen with this corpus. It is unlikely either level of learner would show any difference in vocabulary size if VST-tested before and after this varied diet of texts. (Sequential texts, such as chapters of a book or theme-based selections, typically have far more recurrence across texts than a set of unrelated texts.)



RANGE PROFILE FOR 14 TEXTS
1299 fams, 6,421 tokens

Range	Number fams at this range	% fams at this range	Cumulative percent
range=1	754 fams	58.04	58.04
range=2	222 fams	17.09	75.13
range=3	134 fams	10.32	85.45
range=4	66 fams	5.08	90.53
range=5	42 fams	3.23	93.76
range=6	24 fams	1.85	95.61
range=7	16 fams	1.23	96.84
range=8	9 fams	0.69	97.53
range=9	6 fams	0.46	97.99
range=10	5 fams	0.38	98.37
range=11	3 fams	0.23	98.60
range=12	4 fams	0.31	98.91
range=13	2 fams	0.15	99.06
range=14	12 fams	0.92	99.98

0 items unclassified

Fig. 3. Range profile of the reading corpus

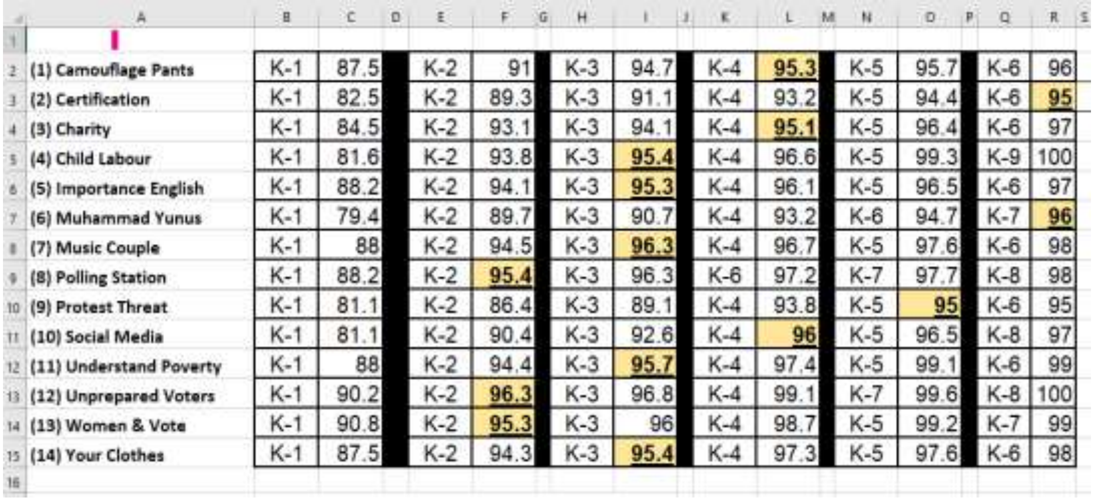
So we see that these materials have a number of limitations. The analysis at this point could proceed to several further aspects of their usability and learnability, but instead turns to data-driven strategies for relieving the problems found in them thus far.

3. Data to the rescue

It would be possible to rewrite these materials substantially to make them match their intended learners' abilities better and offer them better learning opportunities. *Vocabprofile* has an Edit-to-a-Profile facility to help course writers write words up or down, that is, replace particular words with simpler or more complex synonyms, but this is time-consuming work. However, there are a number of simpler things that can be done to add value to the existing materials.

3.1 Sequencing from easier to harder

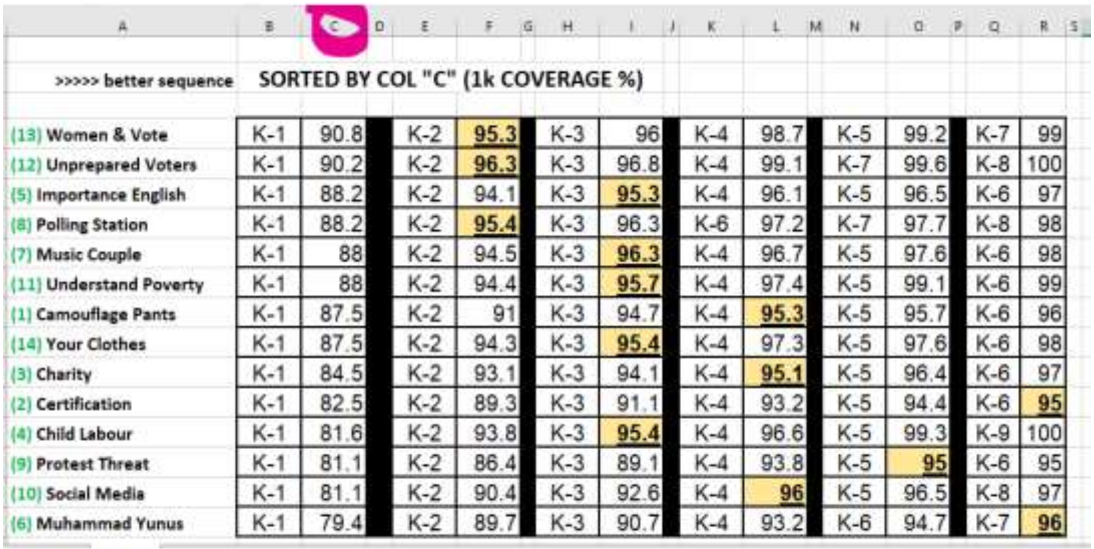
The lexical profile presented in Section 2.1 was for the corpus as a whole, but in fact the profiles differ from text to text. *Vocabprofile* offers a profile summary that can be extracted from the analysis one text at a time and then sorted in Excel or other spreadsheet software. Since the texts are in no particular sequence, they might as well be organized from easier to more difficult – easier in the sense of having a greater component of higher frequency, particularly K-1, words. Figure 4 shows the collected individual profiles of the texts in the corpus with the percentage of words at each K-level from K-1 to K-6 and highlighting the point where 95% (and plausible learnability) is reached. In Text 8 ('Polling Station'), the 95% criterion is reached at K-2; in Text 2 ('Certification'), it is reached at K-6. Like the topics, the lexical sophistication is random.



A	B	C	E	F	H	I	K	L	N	O	Q	R
(1) Camouflage Pants	K-1	87.5	K-2	91	K-3	94.7	K-4	95.3	K-5	95.7	K-6	96
(2) Certification	K-1	82.5	K-2	89.3	K-3	91.1	K-4	93.2	K-5	94.4	K-6	95
(3) Charity	K-1	84.5	K-2	93.1	K-3	94.1	K-4	95.1	K-5	96.4	K-6	97
(4) Child Labour	K-1	81.6	K-2	93.8	K-3	95.4	K-4	96.6	K-5	99.3	K-9	100
(5) Importance English	K-1	88.2	K-2	94.1	K-3	95.3	K-4	96.1	K-5	96.5	K-6	97
(6) Muhammad Yunus	K-1	79.4	K-2	89.7	K-3	90.7	K-4	93.2	K-6	94.7	K-7	96
(7) Music Couple	K-1	88	K-2	94.5	K-3	96.3	K-4	96.7	K-5	97.6	K-6	98
(8) Polling Station	K-1	88.2	K-2	95.4	K-3	96.3	K-6	97.2	K-7	97.7	K-8	98
(9) Protest Threat	K-1	81.1	K-2	86.4	K-3	89.1	K-4	93.8	K-5	95	K-6	95
(10) Social Media	K-1	81.1	K-2	90.4	K-3	92.6	K-4	96	K-5	96.5	K-8	97
(11) Understand Poverty	K-1	88	K-2	94.4	K-3	95.7	K-4	97.4	K-5	99.1	K-6	99
(12) Unprepared Voters	K-1	90.2	K-2	96.3	K-3	96.8	K-4	99.1	K-7	99.6	K-8	100
(13) Women & Vote	K-1	90.8	K-2	95.3	K-3	96	K-4	98.7	K-5	99.2	K-7	99
(14) Your Clothes	K-1	87.5	K-2	94.3	K-3	95.4	K-4	97.3	K-5	97.6	K-6	98

Fig. 4. Random arrivals at the 95% point

With this knowledge, it is a simple matter to re-sort the worksheet by Column “C” (K-1 coverage percentage) to obtain a more coherent learnability sequence, as shown in Figure 5. If desired, a secondary sort could be requested for Column “E” (K-2), and so on, but this was not done here. Once re-sorted, the texts met first will be those with a higher proportion of common vocabulary. Then the numbering can be changed and the materials reassembled. In this case, three of the first four texts encountered meet their 95% criterion with 1,000 word families. This procedure clearly accommodates the weaker learners; it could easily be flipped to accommodate the stronger – starting with the more sophisticated texts, or replacing easier with more difficult texts.



>>>>> better sequence SORTED BY COL "C" (1k COVERAGE %)

A	B	C	E	F	H	I	K	L	N	O	Q	R
(13) Women & Vote	K-1	90.8	K-2	**95.3**	K-3	96	K-4	98.7	K-5	99.2	K-7	99
(12) Unprepared Voters	K-1	90.2	K-2	**96.3**	K-3	96.8	K-4	99.1	K-7	99.6	K-8	100
(5) Importance English	K-1	88.2	K-2	94.1	K-3	**95.3**	K-4	96.1	K-5	96.5	K-6	97
(8) Polling Station	K-1	88.2	K-2	**95.4**	K-3	96.3	K-6	97.2	K-7	97.7	K-8	98
(7) Music Couple	K-1	88	K-2	94.5	K-3	**96.3**	K-4	96.7	K-5	97.6	K-6	98
(11) Understand Poverty	K-1	88	K-2	94.4	K-3	**95.7**	K-4	97.4	K-5	99.1	K-6	99
(1) Camouflage Pants	K-1	87.5	K-2	91	K-3	94.7	K-4	**95.3**	K-5	95.7	K-6	96
(14) Your Clothes	K-1	87.5	K-2	94.3	K-3	**95.4**	K-4	97.3	K-5	97.6	K-6	98
(3) Charity	K-1	84.5	K-2	93.1	K-3	94.1	K-4	**95.1**	K-5	96.4	K-6	97
(2) Certification	K-1	82.5	K-2	89.3	K-3	91.1	K-4	93.2	K-5	94.4	K-6	**95**
(4) Child Labour	K-1	81.6	K-2	93.8	K-3	**95.4**	K-4	96.6	K-5	99.3	K-9	100
(9) Protest Threat	K-1	81.1	K-2	86.4	K-3	89.1	K-4	93.8	K-5	**95**	K-6	95
(10) Social Media	K-1	81.1	K-2	90.4	K-3	92.6	K-4	**96**	K-5	96.5	K-8	97
(6) Muhammad Yunus	K-1	79.4	K-2	89.7	K-3	90.7	K-4	93.2	K-6	94.7	K-7	**96**

Fig. 5. Staged arrivals at the 95% point

3.2 A corpus-based vocabulary supplement

For the very weak learners, however, just meeting words in a more rational sequence will still not be sufficient either for independent reading or further vocabulary growth. These learners need more encounters, such as would be provided by a vocabulary supplement, where new words met in the text are met again in a different or more memorable format or more comprehensible context. From a corpus *Range* can easily assemble a vocabulary list based on both frequency and range, so that words with a high range across the corpus and frequency within the corpus can be extracted for further more targeted work. Figure 6 is a list of 46 words pulled out of the corpus with the following specifications selected in the checkboxes:

- Not a function word
- Has at least 7 corpus occurrences
- Appears in five texts or more (from 14)

able	age	allow	also	become	better
child	come	country	day	demand	economy
even	family	find	force	give	govern
help	important	issue	last	like	live
need	now	order	part	people	poor
problem	public	reason	say	school	service
show	situation	think	time	use	want
way	world	work			

Fig. 6. A basic data-driven word bank

These are the words that can be considered candidates for input-based or incidental learning. Many of the words will already be known to some class members, but with four out of 11 in the sample group (36%) having little vocabulary at all, they can profit from having their attention drawn to common words (*want, show, way, need, think*). The list will make teachers aware of which common words their learners will be meeting in the readings. For more advanced learners, the corpus can be re-run through *Range* excluding not just function words but all K-1 words. This produces the list shown in Figure 7, which, though similar to Figure 6 from a native speaker's perspective, contains a number of interesting challenges for the upper-intermediate Francophone learner. These include words that have no related form in French (*threat, increase, remove*), or have a misleading similarity of form (*union, due, demand*), or present a pronunciation challenge (*develop, economy*). Such a list allows a teacher not just to respond to learners' needs as they arise but to anticipate them. Basic and advanced lists could be used by the teacher or given to the relevant learners themselves, whether as flashcards, example sentences from the course corpus or another corpus, or even as raw lists.

active	advance	create	demand	develop	due
economy	effort	etc	example	extreme	include
increase	individual	industry	interact	labour	organize

politics	population	poverty	provide	province	remove
society	text	threat	union	vote	

Fig. 7. A less basic data-driven word bank from the same corpus

Yet it could be argued that single family headwords (as shown in Figures 7 and 8) without associated family members in their various morphologies are of limited value. Additional forms of the same word can often be learned for little additional effort (Laufer & Cobb 2019) and thus should be available when or shortly after a new word is encountered. Any headword can be fleshed out to its full family at Fami/Lemmatizer (www.lextutor.ca/familizer), which can then be used in, for example, a word bank to appear at the end of the course book. Such a glossary is shown in Figure 8 for the first few items of Figure 7.

active actively activism activist activists activities activity inactive inactivity
advance advanced advancement advances advancing
create created creates creating creation creations creative creatively creativity creator creators recreate recreated recreates recreating
demand demanded demanding demands undemanding

Fig. 8. Headwords fleshed out

However, a complete listing of all possible family members may be more than is required or useful, especially in the case of derived forms that are either infrequent (*active, activism*) or involve substantial change to the base form or pronunciation (*create, creativity*). A better approach to creating a course word bank is to start with a complete vocabulary list and pare it down to a nucleus of just those members that actually appear a certain number of times in the course corpus. *Nuclear List Builder* (www.lextutor.ca/freq/nuclear, Cobb & Laufer, in press) performs exactly this function, reducing Nation's (2016) BNC-COCA lists by K-level to just those forms found in a corpus (of, e.g., the learners' course materials). A way to tailor this concept and software to these learners' needs would be to draw K-1, K-2, and K-3 nuclear lists out of the full lists and match them to the three populations of learners identified by vocabulary testing. Part of the K-1 version of this list is shown in Figure 9a, K-3 in Figure 9b.

These tailored nuclear lists are clearly reduced in both number and size of families, compared to the original lists. Both lists originally comprised 1,000 families, as the ‘K’ in their names indicates, or 6,862 words including all family members at K-1 and 5,884 members at K-3. The nuclearized K-1 list is 369 families with 479 members; K-3 is 60 families with 68 members.

A further advantage to nuclear lists is that in subsequent reading courses for the same learners, new and probably somewhat overlapping lists could be drawn from their new learning materials, such that the vocabulary component was spiraling through new and familiar items and eventually covering a substantial number of them in a substantial number of occurrences and contexts.

- a
  - an
- able
  - ability
- about
- across
- act
  - action
- actually
- advertisements
- afford
- after
- again
- against
- age
  - ages
  - underage
- ago
- agree
- ahead
- all
- allowance
  - allowed
- almost
- along
- also
- although
- always
- amount

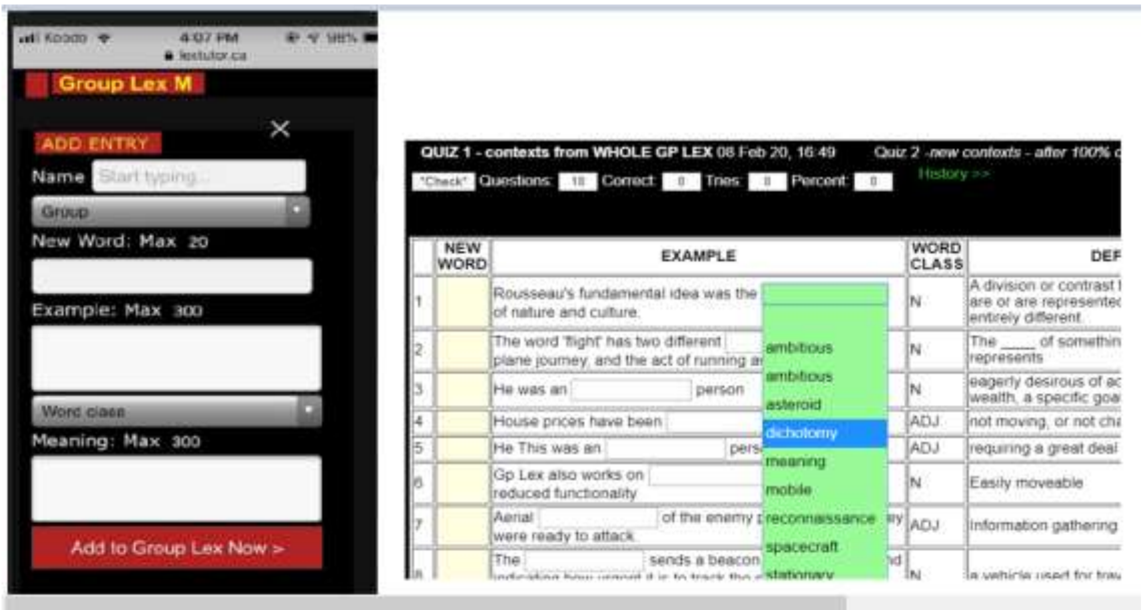
Fig. 9a. Part of BNC-COCA K-1 Nuclear

- accomplished
- addiction
- authority
- awards
- charity
- client
  - clients
- communicate
  - communication
- concert

consumer
content
controversy
controversial
controversies

Fig. 9b. Part of BNC-COCA K-3 Nuclear

Once created, such lists could be used in a number of ways. They might be provided just to teachers, who could then anticipate which words would be most usefully emphasized in discussing the texts or constructing worksheets, or they could be given to students as the basis of a learning activity to supplement input-based learning. For the latter, one approach would be to choose one of the data-driven learning tools featured in Boulton & Cobb's (2017) meta-analysis of DDL tools, such as Lextutor's *Group Lex* (www.lextutor.ca/group_lex). With this routine, learners can enter, say, 10 words per week from their list into an app on their computers, pads, or mobile phones, accompanied by an example and a brief meaning, and then quiz themselves on their own and their classmates' entries as they effectively co-construct their own K-1 and K-3 level-appropriate lexicons or dictionaries. Or the teacher can use the software to create whole-class paper quizzes. The phone input and a full-size sample quiz are shown in Figure 10.



QUIZ 1 - contexts from WHOLE GP LEX 08 Feb 20, 16:49 Quiz 2 -new contexts - after 100% c

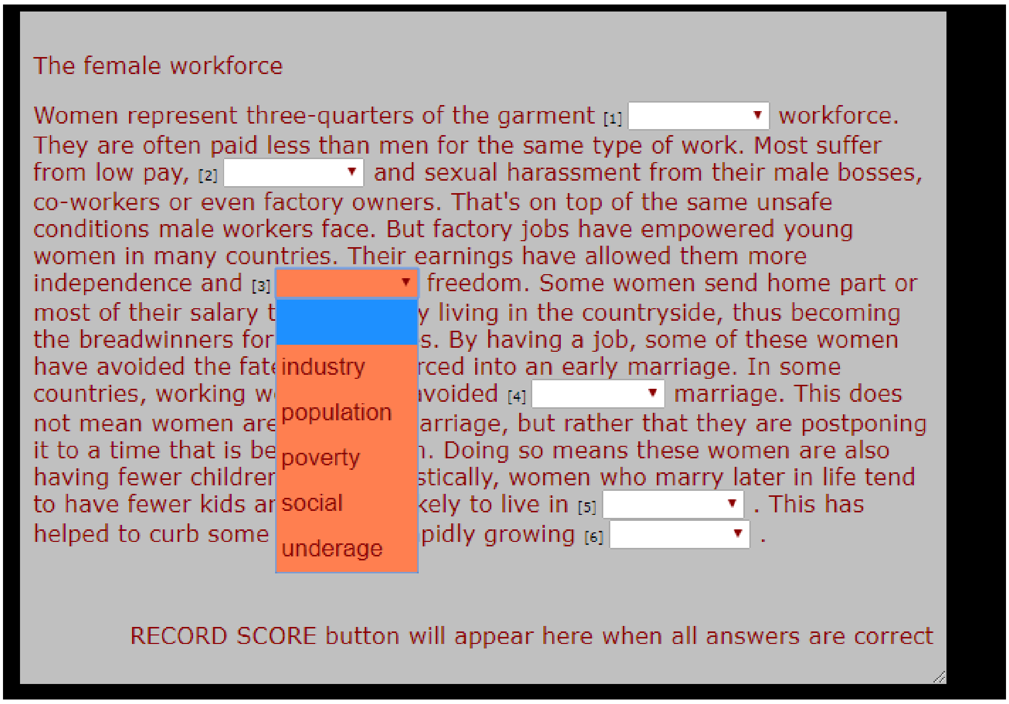
"Check" Questions: 10 Correct: 0 Tries: 0 Percent: 0 History >>

	NEW WORD	EXAMPLE	WORD CLASS	DEF
1		Rousseau's fundamental idea was the of nature and culture.	N	A division or contrast t are or are representec entirely different.
2		The word 'flight' has two different plane journey, and the act of running a	N	The _____ of somethin represents
3		He was an person	N	eagerly desirous of ac wealth, a specific goa
4		House prices have been	ADJ	not moving, or not ch
5		He This was an pers	ADJ	requiring a great deal
6		Gp Lex also works on reduced functionality	N	Easily moveable
7		Aerial of the enemy p ey were ready to attack.	ADJ	Information gathering
8		The sends a beacon nd	N	a vehicle used for trav

Fig. 10. Lists to lexicons

Another focus on key words as they actually appear in context in the corpus would be to put lexically rich portions of course texts into Lextutor's

Clozebuilder (www.lextutor.ca/cgi-bin/cloze/n/) producing an online or paper activity like the one shown in Figure 11 from the course text 'Your Clothes.' The teacher can select which words to focus on, by principle instead of by guessing, and can add extra resources like dictionary or text-to-speech.



The female workforce

Women represent three-quarters of the garment [1] ▾ workforce.
They are often paid less than men for the same type of work. Most suffer
from low pay, [2] ▾ and sexual harassment from their male bosses,
co-workers or even factory owners. That's on top of the same unsafe
conditions male workers face. But factory jobs have empowered young
women in many countries. Their earnings have allowed them more
independence and [3] ▾ freedom. Some women send home part or
most of their salary t y living in the countryside, thus becoming
the breadwinners for s. By having a job, some of these women
have avoided the fat rced into an early marriage. In some
countries, working w voided [4] ▾ marriage. This does
not mean women are arriage, but rather that they are postponing
it to a time that is be n. Doing so means these women are also
having fewer childrer stically, women who marry later in life tend
to have fewer kids ar kely to live in [5] ▾ . This has
helped to curb some pidly growing [6] ▾ .

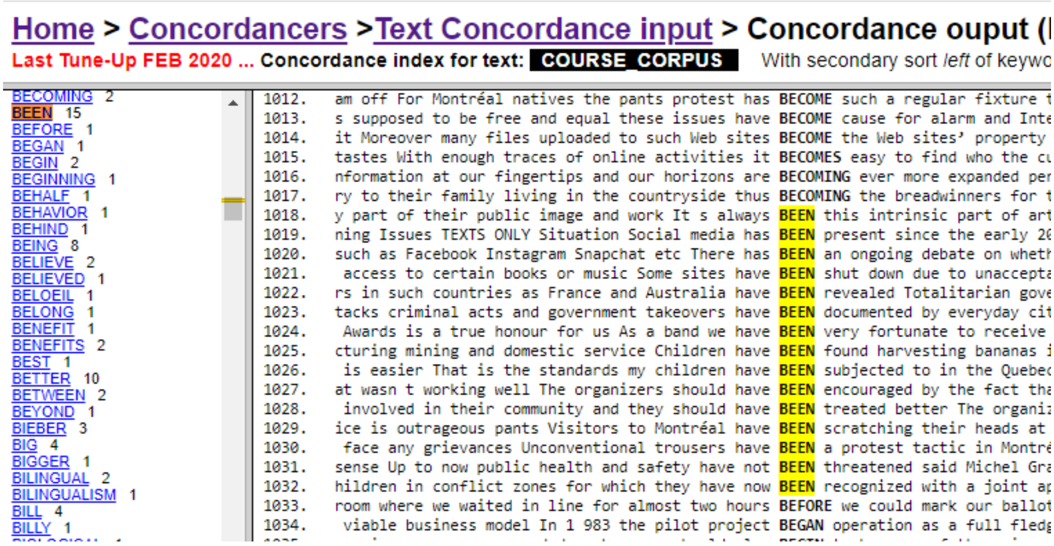
RECORD SCORE button will appear here when all answers are correct

Fig. 11. List words in context

4. Adding a grammar component

While vocabulary knowledge is the main component of reading comprehension, comprehension can also be limited by grammar knowledge. It is worth knowing what grammatical features are present or predominant in the course corpus and planning some course time around them. As is well known, different text types have their own grammatical profiles: news stories often rely on past tense, opinion pieces on present tense, scientific pieces on passive voice, etc. To determine the grammar profile of the corpus, it can be entered as a single file into a Lextutor's *Text Concordance* (www.lextutor.ca/conc/text/).

This program provides a teacher or course designer with a grammatical snapshot of the corpus as a whole. Frequent words from the listing on the left in Figure 12a can be clicked to reveal the type of environment each comes from, grammatical as well as lexical. Clicking *been* for example leads to 15 instances of present perfect verbs, of which nine are passives and one continuous, suggesting that both present perfect and passivization could at some point be worth reviewing or anticipating as a source of miscomprehension.



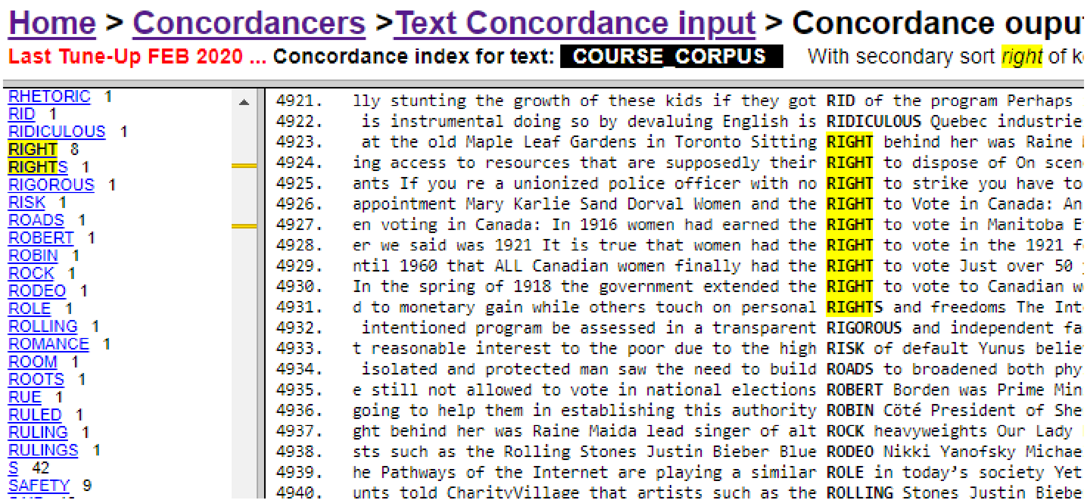
Home > Concordancers >Text Concordance input > Concordance ouput (

Last Tune-Up FEB 2020 ... Concordance index for text: COURSE_CORPUS With secondary sort *left* of keywo

```
BECOMING  2
BEEN  15
BEFORE  1
BEGAN  1
BEGIN  2
BEGINNING  1
BEHALF  1
BEHAVIOR  1
BEHIND  1
BEING  8
BELIEVE  2
BELIEVED  1
BELOEIL  1
BELONG  1
BENEFIT  1
BENEFITS  2
BEST  1
BETTER  10
BETWEEN  2
BEYOND  1
BIEBER  3
BIG  4
BIGGER  1
BILINGUAL  2
BILINGUALISM  1
BILL  4
BILLY  1
```

```
1012.   am off For Montréal natives the pants protest has BECOME such a regular fixture t
1013.   s supposed to be free and equal these issues have BECOME cause for alarm and Inte
1014.   it Moreover many files uploaded to such Web sites BECOME the Web sites’ property
1015.   tastes With enough traces of online activities it BECOMES easy to find who the cu
1016.   nformation at our fingertips and our horizons are BECOMING ever more expanded per
1017.   ry to their family living in the countryside thus BECOMING the breadwinners for t
1018.   y part of their public image and work It s always BEEN this intrinsic part of art
1019.   ning Issues TEXTS ONLY Situation Social media has BEEN present since the early 20
1020.   such as Facebook Instagram Snapchat etc There has BEEN an ongoing debate on wheth
1021.    access to certain books or music Some sites have BEEN shut down due to unaccepta
1022.   rs in such countries as France and Australia have BEEN revealed Totalitarian gove
1023.   tacks criminal acts and government takeovers have BEEN documented by everyday cit
1024.    Awards is a true honour for us As a band we have BEEN very fortunate to receive
1025.   cturing mining and domestic service Children have BEEN found harvesting bananas i
1026.    is easier That is the standards my children have BEEN subjected to in the Quebec
1027.   at wasn t working well The organizers should have BEEN encouraged by the fact tha
1028.    involved in their community and they should have BEEN treated better The organiz
1029.   ice is outrageous pants Visitors to Montréal have BEEN scratching their heads at
1030.    face any grievances Unconventional trousers have BEEN a protest tactic in Montré
1031.   sense Up to now public health and safety have not BEEN threatened said Michel Gra
1032.   hildren in conflict zones for which they have now BEEN recognized with a joint ap
1033.   room where we waited in line for almost two hours BEFORE we could mark our ballot
1034.    viable business model In 1 983 the pilot project BEGAN operation as a full fledg
```

Fig. 12a. Grammar in the course corpus



Home > Concordancers >Text Concordance input > Concordance oupu

Last Tune-Up FEB 2020 ... Concordance index for text: COURSE_CORPUS With secondary sort *right* of k

```
RHETORIC  1
RID  1
RIDICULOUS  1
RIGHT  8
RIGHTS  1
RIGOROUS  1
RISK  1
ROADS  1
ROBERT  1
ROBIN  1
ROCK  1
RODEO  1
ROLE  1
ROLLING  1
ROMANCE  1
ROOM  1
ROOTS  1
RUE  1
RULED  1
RULING  1
RULINGS  1
S  42
SAFETY  9
```

```
4921.   lly stunting the growth of these kids if they got RID of the program Perhaps
4922.    is instrumental doing so by devaluing English is RIDICULOUS Quebec industrie
4923.    at the old Maple Leaf Gardens in Toronto Sitting RIGHT behind her was Raine
4924.   ing access to resources that are supposedly their RIGHT to dispose of On scen
4925.   ants If you re a unionized police officer with no RIGHT to strike you have to
4926.   appointment Mary Karlie Sand Dorval Women and the RIGHT to Vote in Canada: An
4927.   en voting in Canada: In 1916 women had earned the RIGHT to vote in Manitoba E
4928.   er we said was 1921 It is true that women had the RIGHT to vote in the 1921 f
4929.   ntil 1960 that ALL Canadian women finally had the RIGHT to vote Just over 50
4930.   In the spring of 1918 the government extended the RIGHT to vote to Canadian w
4931.   d to monetary gain while others touch on personal RIGHTS and freedoms The Int
4932.    intentioned program be assessed in a transparent RIGOROUS and independent fa
4933.   t reasonable interest to the poor due to the high RISK of default Yunus belie
4934.    isolated and protected man saw the need to build ROADS to broadened both phy
4935.   e still not allowed to vote in national elections ROBERT Borden was Prime Min
4936.   going to help them in establishing this authority ROBIN Côté President of She
4937.   ght behind her was Raine Maida lead singer of alt ROCK heavyweights Our Lady
4938.   sts such as the Rolling Stones Justin Bieber Blue RODEO Nikki Yanofsky Michae
4939.   he Pathways of the Internet are playing a similar ROLE in today’s society Yet
4940.   unts told CharityVillage that artists such as the ROLLING Stones Justin Biebe
```

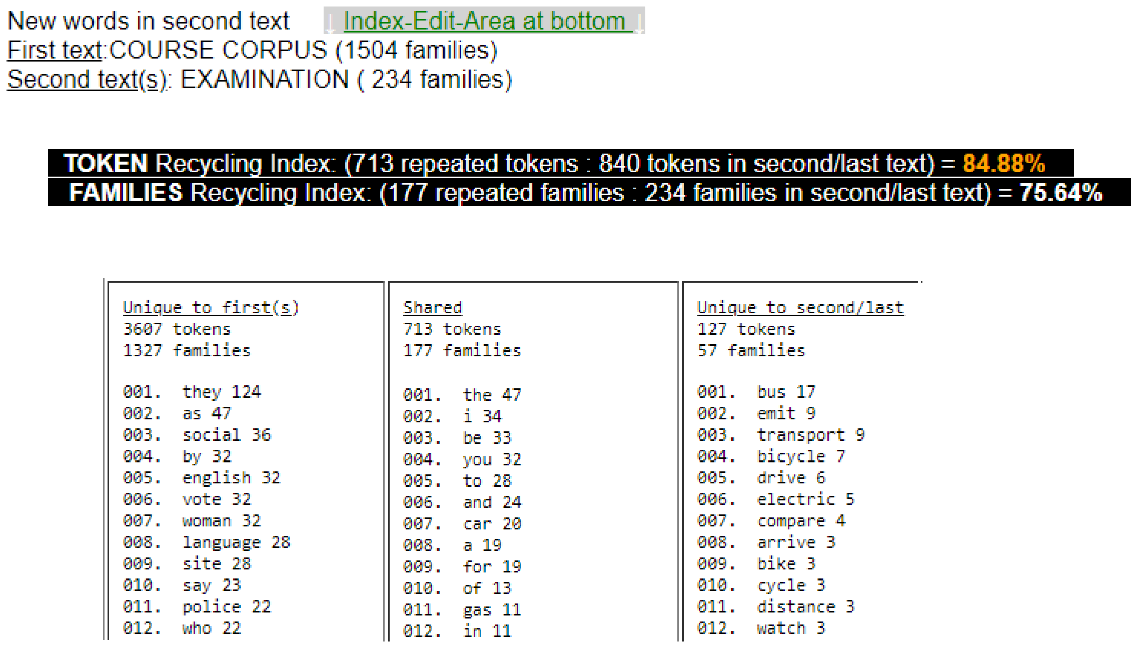
Fig. 12b. Collocation in the course corpus

Any recurring collocation pattern can also be easily identified. Figure 12b shows that the polysemous word *right* appears in the corpus only in the context of *the right to* do something, but that it never appears in the contexts of *right and left* or *right and wrong*. However, a teacher who was bilingual would probably notice that constructions employing this word are subtly different from ostensible equivalents in French. *Have a right to* in English tends to be followed by a verb, while the similar-looking *avoir droit à* is invariably followed by a noun; for a verb the expression is *avoir le droit de (partir)*, while in English *the right of* prepares for a noun (like *passage* or *way*). *Text Concordance* alerts a teacher of Francophones to think about a worksheet on this.

4. Examination control

The final use for the course corpus that we will discuss here is to assure that the examination text, usually a new text, and the questions about it line up with what the learners have actually read in the course or could reasonably be expected to have inferred from what they have read. Such an assurance is what Biggs (2014) calls 'constructive alignment.' For a language course at this level, alignment means the test should contain only the words and structures the learners had actually encountered in the course, plus some small space for inference if that was part of the training program. Providing such assurance is not only a pedagogical but also an ethical issue. It applies not only to language examinations but to any that involve reading comprehension (like story problems in mathematics or science).

Alignment of language courses and tests, even on the level of words and phrases, is often not achieved and indeed is not simple to achieve without computation. A routine called *Text Lex Compare* (www.lextutor.ca/cgi-bin/tl_compare/) compares all the words or phrases of an examination text to the collected course texts the learners have worked with. Figure 13 shows part of a comparison between the course corpus under discussion and a passage that was actually used as the basis for an examination in an earlier run of the course. The output consists of the words that are unique to the course (left column), the words that are shared between course and examination text (middle column), and the words that are unique to the examination text (right column). As shown in the heading information, just under 15% of the word tokens (individual words) or 24% of word families appearing on the test had not previously been met in the course at all. This is a far higher proportion than learners can be expected to work out making inferences from context (native speakers can manage 2%; Schmitt et al. 2011).



New words in second text Index-Edit-Area at bottom
First text:COURSE CORPUS (1504 families)
Second text(s): EXAMINATION (234 families)

TOKEN Recycling Index: (713 repeated tokens : 840 tokens in second/last text) = **84.88%**
FAMILIES Recycling Index: (177 repeated families : 234 families in second/last text) = **75.64%**

Unique to first(s)	Shared	Unique to second/last
3607 tokens	713 tokens	127 tokens
1327 families	177 families	57 families
001. they 124	001. the 47	001. bus 17
002. as 47	002. i 34	002. emit 9
003. social 36	003. be 33	003. transport 9
004. by 32	004. you 32	004. bicycle 7
005. english 32	005. to 28	005. drive 6
006. vote 32	006. and 24	006. electric 5
007. woman 32	007. car 20	007. compare 4
008. language 28	008. a 19	008. arrive 3
009. site 28	009. for 19	009. bike 3
010. say 23	010. of 13	010. cycle 3
011. police 22	011. gas 11	011. distance 3
012. who 22	012. in 11	012. watch 3

Fig. 13. Data-driven test control

From here three remedial interventions are possible. One is to re-write the examination text, focusing on the words in Column 3, either deciding these are acceptable as known cognates with French (*bus, transport, bicycle*) or else glossing/elaborating their meanings inside the text or in a supplementary glossary for non-cognate items (*drive, watch, bike*). Another possibility is to leave 2% of non-cognate items for contextual inference, assuming this has been a focus of the course and is thus testable.

4. Conclusion

The main point of interest in this sequence of DDCD measures is how few of the patterns discussed would have been observed without the data aggregation of the corpus and the computational tools to look at it. There is too much information in a set of texts to be gleaned by the naked eye; a technology is required to extend the pedagogical sensorium.

The hypothesis of the paper was that a reading couse of online texts can be usefully treated as a corpus and investigated then improved with text analysis tools. Readers can judge whether the argument has been made, but for the writer it seems clear that following the procedures outlined above will produce a more successfu reading course. This proposition would have to be tested empirically, in the context of its integration in a real course, though many of the procedures have been tested in smaller and more controlled contexts already. Well established research findings include these: Reading instruction is unsuccessful with less than 90% word knowledge; vocabulary growth is weak without numerous encounters with words; collocational *faux-amis* can impede comprehension; assessment is regularly misaligned with what was taught. These are separate research insights, however, and are only potentially useful until they are integrated into what learners actually do over extended periods of instruction - in other words, into courses. Such an integration is extremely difficult to achieve by individual teachers with a handful of random texts and a roomful of roughly placed learners; it can only be done on a system-wide scale with the aid of instrumentation. Validated instruments are needed to test and place the learners, analyse and sequence the materials, scan materials for obstacles, and correct or supplement them in line with what is known about learning. The particular sequence of tool deployment modeled here – build corpus, profile vocabulary, test learners, re-sequence materials, scan for grammar and collocations, supplement materials, test in line with course characteristics - is probably the most plausible sequence, though others are possible and the software is set up to encourage experimentation.

Would such a redesigned course lead to significant improvements in achievement over the course as originally conceived? It is plausible that it would, because the course would be built from pieces that have already been validated individually. But the pieces would have to be empirically validated again in this larger context, and doing this involves a number of challenges. Course comparison research or any large-scale comparison is rarely attempted in applied linguistics, or education research generally, because of the number of variables, the extended time frame, the potential interactions, and the likelihood of confounding. Nevertheless, it is arguably time to expand course evaluation methodologies, which measure courses against objectives, into course comparison methodologies. Most applied linguists would probably agree that research and implementation are out of balance. Without course/curriculum comparisons, large scale implementations of research findings will be slower than they need to be or may never happen at all. The challenge is to maintain experimental rigour while scaling up to real-world settings.

To summarize, evidence-based pedagogy has been difficult to implement at the course or curriculum level, despite the wealth of single-study research results, whether in vocabulary or other areas of language learning. Vocabulary just makes a particularly clear case. Between the research and the learner lies a void that has seemed difficult to fill. Data-driven course design is a strong contender to fill it.

REFERENCES

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

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics 25* (1), 133-150.

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education 1*, 5-22.

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning 65* (2), 1-46.

British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at: http://www.natcorp.ox.ac.uk/.

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 & Technology 11* (3), 38-63.

Cobb, T. & Laufer, B. (In press.) A Nuclear word family list: A list of most frequent family members - base words and derived words. *Language Learning 71* (3).

Crossley, S., Kyle, K. & McNamara. D. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods 48* (4), 1227-1237.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present.* Available online at https://www.english-corpora.org/coca/.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal 20* (3), 465-480.

Greaves, C. (2009). *Concgram 1.0: A phraseological search engine.* Amsterdam: John Benjamins.

Hu, M. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language 13* (1), 403-430.

Laufer, B. & Cobb, T. (2019). How much knowledge of derived words is needed for reading? *Applied Linguistics,* November. https://doi.org/10.1093/applin/amz051

Laufer, B. & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learner's vocabulary size and reading comprehension, *Reading in a Foreign Language, 22,*15–30.

Lee, H., Warschauer, M. & Lee, J. (2020). Toward the Establishment of a Data-Driven Learning Model: Role of Learner Factors in Corpus-Based Second Language Vocabulary Learning. *Modern Language Journal 104,* 2. https://doi.org/10.1111/modl.12634

Nation, P. (2016). *Making and using word lists for language learning and testing.* Amsterdam, Netherlands: Benjamins. https://doi.org/10.1075/z.208

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

Schmitt, N., Jiang, X. & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal, 95,* 26-43.

Webb, S. & Nation, P. (2017). *How vocabulary is learned.* New York: Oxford University Press.