The Encyclopedia of Applied Linguistics

Technology and Learning Vocabulary

Journal:	*The Encyclopedia of Applied Linguistics*
Manuscript ID:	AL-2010-0634.R1
Wiley - Manuscript type:	article
Date Submitted by the Author:	13-Nov-2010
Complete List of Authors:	Cobb, Thomas; Université du Québec à Montréal, Didactique des langues
Keywords:	literacy, language learning technology, Educational Linguistics, Second Language Acquisition, vocabulary, writing

Technology and vocabulary learning

Thomas Cobb

Université du Québec à Montréal

cobb.tom@uqam.ca

Word count: 3659

Reference word count: 663

Key words : learning, lexis, media, corpus, lexicon, speech, writing, vocabulary.

To date, vocabulary is the dimension of language that has been most intimately involved with technology. Vocabulary is the subject of more articles in this section of the Encyclopedia than phonology, syntax, or pragmatics, and it has been subject to several distinct types of computational analysis – lexicographical, terminological, collocational, keyword, frequency, and others. But in terms of vocabulary learning, a simpler and older technology than any of these has had and continues to have the greatest effect on the kind and number of words that people can learn, and that is the technology of writing (Ong, 1982). When the practice of writing words down becomes widespread in a language, it both determines the vocabulary that will have to be to be learned as well as offering the means for learning it.

That said, much of the vocabulary that anyone knows has been learned without any involvement of a technology at all. Infants hear speech streams, gradually infer their global meanings from context, and more gradually still identify their units, including words. But lexicons acquired exclusively in this manner without the assistance of writing are destined to remain rudimentary, whether on the level of individuals or communities. Writing, as a recoding of the speech stream that slows its rate and enables repetition, review, and parsing, promotes far more extensive learning than real-time language use alone.

But as with many technologies, writing comes with a downside. The effect of making lexicons easier to learn by writing them down has been to make the overall task of vocabulary learning more difficult, because writing allows lexicons to expand up to and beyond the point of learnability. This article looks at how writing and its allied technologies first made lexicons easier to learn, then harder to learn, but are now making them easy again.

[A] The vocabulary to be learned

When a language is spoken but not written, as was the case with all languages until relatively recently, and remains the case with most (Coulmas, 1989), there is a limit on how large its lexicon can become, one presumably imposed by the human memory and lifespan. If users of unwritten languages are required to encode progressively more complex messages, they tend to carve further distinctions within words (declensions, case

endings, and verb categories) rather than coining new words, probably because it is easier to learn variations on known items than new items. Without a written form to suggest what the words in the speech stream actually are – the pieces of language between empty spaces – it is not clear the speech stream is ever very clearly parsed. Several unwritten languages "do not have a word for *word*" (Coulmas, 1989, p. 20). Research with language users that have for some reason missed out on acquiring literacy suggests that they may not have a word for *word* either. Tarone studied the language use of young Somali immigrants to the United States who had never learned a written language, and found they lacked very basic word handling skills such as repeating pseudowords and building associations between phonologically similar units. Such inabilities can be predicted to limit vocabulary learning from oral input.

Vocabulary size counts of preliterate languages are necessarily inferential, but it seems that about 10,000 lemmas (head words plus inflections and tenses) is the norm. This is the calculated size of several unwritten Australian languages (Dixon, 1980, p.2). Even in the 12th century, five centuries after Old English was given a written form by Christian missionaries, it comprised only 20,000 lemmas (based on a count of the University of Toronto Corpus). English now comprises some 470 thousand headwords (in Webster's Third New International Dictionary, Unabridged, with the 1993 Addenda). English is always a special case, with its many incoming and outgoing linguistic invasions, but the pattern obtains: there is no major written language with a lexicon as small as 10,000 words. Of course, no current speaker of English knows anything like all of its lexicon. Goulden, Nation and Read (1990) calculated that even educated adult native speakers have recognition knowledge of only 13,000 to 17,000 word families, so it is reasonable to infer that less educated or semi-literate adults would have knowledge of fewer words than that, or something roughly comparable to the size of an unwritten language.

When words are written down in a consistent form with spaces on either side, some of the burden of identifying and remembering them is offloaded to an external memory, and hence more words can be tracked and accessed. (End note 1). Words written down are also more difficult for the language to prune through collective forgetting or dark ages; written lexicons tend to pile up relentlessly. Of course, not all ways of writing words down lift the memory burdens of vocabulary learning equally. Writing systems are constituted to favour one of two constituencies, those learning a language or its experienced users (Sampson, 1985; Venezky, 2004). Spanish with its abundant and reliable phonological information and small number of perceptual units (letters and letter parts) favours the learner; words known in speech can easily be identified in writing, and vice-versa. Chinese with its minimal phonological information and vast number of perceptual units favours rapid processing by those who manage to learn it. A logographic or Hanzi text can be read about three times faster than the same text written in romanized Pinyin by readers who are equally experienced in both scripts (according to Sun & Feng, 1999).

Very large lexicons generally, in whatever writing system, also favour competent users at the expense of learners. There are simply a lot of words to learn in a language like English or French. Academic underachievement is a topic of growing concern in

educational policy and research in most Western countries, and increasingly researchers are identifying lexical inadequacy as a major cause. Cummins (e.g., 1981) observed that children often master the vocabulary of interpersonal communication but fail to develop the "cognitive and academic" language proficiency that is encoded in less familiar lexis. Corson (1985) went further postulating a "lexical bar" consisting of the terms of science, analysis and research which many children never become familiar with. Chall and Jacobs (2003) identified a widespread "fourth grade slump" at about age 10 when school children move beyond the lexis of familiar situations toward that of written texts.

The problem of inadequate lexis is particularly acute for those who wish to function at a high level in a second language (L2), as a large and growing number presently wish to do, especially in English. Perhaps this explains why it was mainly researchers concerned with English as an L2 who tackled the too-much-vocabulary problem with a view to finding solutions. A number of L2 pedagogical theorists had long experimented with ways of making the vocabulary learning task more manageable for learners of English (West, 1926; Palmer, 1917), but this goal awaited the development of text computing to become truly attainable.

The computational analysis of text is a natural development of the basic technology of writing, a secondary literacy, in terms of what it reveals to language users about the language they are using. Writing systems exposed basic patterns in language, such as where words start and stop, and to this text analysis adds how many and which words there are in a text, in which other texts they also occur, and how frequently or infrequently. It also refines definitions of where words really start and stop, i.e. not necessarily at the spaces, for example in the case of multiword units like "by and large" or "as a matter of fact." But the effects of these two literacies are potentially quite different. If primary literacy expanded the lexicon out of the reach of many L1 and most L2 learners, the secondary literacy can be used to bring it back (Webb & Nation, this volume; Gardner, this volume).

There may be a half million words in English, but how many of them are needed to function in a particular context? Corpus linguists since the 1970s have built progressively more comprehensive models of languages as machine readable text corpora (Kučera & Francis, 1967; Leech et al, 2001). Pedagogical scholars have found ways to use these corpora to calculate the effect of knowing different amounts of vocabulary for different purposes (Nation 2006; Cobb, 2007). Text coverage has now been calculated for several different functional domains including conversation (Adolphs & Schmitt, 2003), academic reading (Coxhead, 2000), domain specific reading (Chung & Nation, 2004), and the discourse of TV and movies (Webb & Rodgers, 2009). Coverage data twinned with comprehension data (e.g., by Schmitt, Jiang, & Grabe, in press) shows reasonably clearly that the 6,000-8,000 most frequent word families are needed for reading in an academic environment, a challenging yet learnable figure. This research is in progress, with several important refinements under way, including Alvarez and Schmitt's (in review) calculation of the coverage of the most frequent multiword units, and Cobb's (in review 1) calculation of the coverage of the separate words buried in high-frequency homographs like "bank." Basic versions of this technology are available for teachers

now, such as the Vocabprofile website (http://lextutor.ca/vp/bnc/) which helps teachers modify learning input to a particular frequency zone. Ashkenazy et al (this volume) are experimenting with Web searches for texts at particular lexical levels. The goal is to help learners focus on the words they need rather than casting them adrift on the open seas of lexis.

From this perspective, then, while one language technology (writing) allowed lexicons to expand to or beyond the limits of learnability, another (text analysis) is busy discovering zones of learnability within these lexicons. But even when clearly identified, such zones will still number word families in the hundreds or even thousands and mastering them will remain a daunting task inevitably left largely to individuals to accomplish. What resources will be available to individuals wishing to undertake a substantial vocabulary upgrade?

[A] The means of learning vocabulary

Learning to read is clearly the first and biggest step in building a lexicon, but is reading alone, even in a targeted lexical zone, the most efficient method of vocabulary expansion? This is one of the most vexed questions in applied linguistics research. A conclusion reached by many researchers (e.g., Cobb, 2008; Laufer, 2005) is that while reading is typically an adequate resource for L1 lexicon building (but see Chall & Jacobs, 2003), it typically is not for L2 lexicon building (but see McQuillan & Krashen, 2008). This section looks at vocabulary learning technologies that go beyond simply meeting words in reading.

Writing is no longer the only technology that can slow language down and render it plain for retention and learning purposes. Numerous types of spoken language media can now perform some version of one or more of these functions, including cassettes, MP3 sound files, and DVD videos. Some media can do this in conjunction with writing, whether pre-recorded audio-texts, videos with subtitling, or real-time text-to-speech created by computer programs. While there is as yet no true text-to-image technology, vast databanks of images both still and moving have been constructed matching much of the high frequency lexicons of several malnguages (e.g., by Rosetta Stone, http://www.rosettastone.com/).

And yet the most interesting approaches to accelerated vocabulary acquisition continue to proceed from written representations with any additional media playing a supplementary role. What the more recent learning technologies add to the basic technology of writing is decontextualized simulations of word handling tasks. These tasks are basically four in number – retrieving forms from meanings, meanings from forms, holding forms in memory, and adapting meanings to novel contexts. A set of word cards remains the basic paradigm for achieving the first three of these.

The data structure of a word card – more likely nowadays to appear on a computer screen or game console that on the two sides of an actual card – divides the word's information into two parts, front and back, shown and hidden, with roles reversible. One part is the to-

be-learned word or phrase, the other some information associated with the word, whether a short definition in L1 or L2, a picture, a translation, a context sentence with the word missing, or other pertinent associate. The learner shuffles the word cards (or clicks Randomize), looks at the cards one at a time, tries to recall the information on the other side, confirms his or her response and undertakes remedial learning as needed, and so on through pack or stack, then repeats. The "other side" can of course be either side. In one round the learner looks at the word and tries to recall the meaning, and in the next looks at the meaning and tries to produce the word. These receptive and productive activities are two of the basic word handling operations mentioned above, and the third – retaining the form in memory – is involved in both of them. An array of strong paired-associate studies from the verbal learning era of psychology (up to roughly 1970) underpins the use of word learning cards, many of them brought to bear on vocabulary expansion in an interesting paper by Mondria and Mondria-De Vries (1994). These researchers propose a plan for recycling the word cards in a shoe box of five compartments: all words start in the first compartment and move toward the fifth and thence out of the system if their information is consistently recalled, or back toward the front if not. Words that need the practice get it, and vice-versa, and numerous words can pass through the system efficiently.

Word cards are also supported by research findings more recent than those of the paired-associate studies. The emphasis on form in word cards, for example, exploits the findings of Barcroft (2006) that the learning of form not meaning is in some ways the greater challenge of word learning. The emphasis on speed and repetitive practice exploits Segalowitz and Hulstijn's (2005) finding that lexical access is key to language processing. Word cards have also proven a useful laboratory for exploring second language learning variables, for example Elgort's (2007) finding that the gloss on the card back is more useful for advanced learners in the target language than in the L1.

Numerous versions of word cards and flashcards now appear on both Internet and mobile devices bringing with them such computational advantages as algorithms for sequencing the presentation, record keeping, and the capacity to make and store large numbers of cards easily. In the case of electronic flash cards, card-back information can include multiple choice items instead of simple translations or glosses, or other word associations and representations in the form of images, sound files, video clips, and whatever other information might be linked from the World Wide Web (iKnow! at http://smart.fm/ allows learners to import Flickr images from www.Flickr.com). Reviews and comparisons of word card learning systems including the following are provided by Godwin-Jones (2010) and Nakata (in press): SuperMemo (www.supermemo.com/), vTrain (www.vtrain.net/), MemoryLifter (www.memorylifter.com/), P-Study (www.takke.jp/), WordChamp (www.wordchamp.com/), Quizlet (quizlet.com/), eSpindle (www.LearnThatWord.org/), and Word Engine (http://www.wordengine.jp/).

However, as Nakata notes few if any of these card systems offer opportunities for generative use of learned items (meeting or using words in new contexts, the fourth basic handling operation mentioned above). And few if any access information from a corpus or dictionary. A system attempting to do some of these is Horst, Cobb and Nicolae's

(2005) *Group Lex* (http://www.lextutor.ca/group_lex/demo/), where learners enter their own words and definitions, hear their words rendered in text-to-speech, build and take quizzes using their own or each other's contexts (Which word goes in this context? Which word corresponds to this gloss?) and then novel contexts from a level-appropriate corpus.

Nor did any of the learning systems reviewed by Nakata appear to (1) focus on timed interactions and automaticity development, (2) tell learners which words they should be looking at beyond those they have found for themselves, or (3) facilitate game interactions or other opportunities for collaborative learning, a key idea in both the socio-constructivist and social networking agendas. One system attempting to do all three is *My Word Coach* (2007) for Nintendo (http://mywordcoach.us.ubi.com/). The tutor inside this set of interactive real-time games knows in advance the entire lexicon of English (or at least the 14,500 most frequent words of the British National Corpus frequency list, Leech et al, 2001), knows the learner, uses a high quality learner dictionary for its glosses (Cambridge Advanced Learner's), encourages wireless collaboration and timed competition between learners, and has undergone extensive testing with strong results on several dimensions of word knowledge (Cobb, in review 2).

[A] Beyond word cards

The use of word cards is of course not the only vocabulary learning technology imaginable. Indeed, no amount of beefing up of the paired associate model can move it very far from the beginner end of the word learning process. Word card systems, pairing words with small glosses or translations, or practising the recognition of form, are closer to the initial phase of word acquisition than the final, that being the complex, generative fitting of known words to novel contexts as required for the productive skills of speech and writing or even the receptive skills of high level reading and real-time speech processing.

Two kinds of technology support for the refinement of existing vocabulary knowledge are available to both L2 and L1 learners. The first is simply the vast array of learning opportunities present in the communications technologies currently available. A learner can receive input in almost any L2 of almost any quantity and quality from any location through devices with options that serve language learning purposes whether or not intended to do so. Media players can be paused and repeated. Radio broadcasts that were interrupted or poorly comprehended can be retrieved as podcasts and listened to again. Hemingway's strategy of learning French by reading the daily news first in English and then in French was only possible because he lived in Paris; but now with newspapers in every major and most minor languages online, any learner can adopt this practice for any pair of languages for the price of an Internet connection. Language learners without their own corpora and concordance programs can Google for their collocates in a writing assignment (Is it "approaches to a solution" or "approaches for"?). The opportunities for vocabulary enrichment in integrated gaming environments like *Second Life* have barely been contemplated (see Milton, this volume).

But there are some more limited, deliberate technologies for deepening existing vocabulary knowledge. A number of these at various degrees of trialing and testing are under development on the author's Lextutor website to support the more advanced learner's vocabulary development in contexts of actual language use. This stage of vocabulary learning may be better named "lexicogrammatical learning" since grammar becomes more relevant as a feature of word use as learning proceeds. Some examples follow.

[B] Reading: Lextutor's hypertext builders link users' texts to a set of click-on resources including text-to-speech for individual words or short phrases, recorded speech or video where available for whole texts, dictionary or concordances for any word or expression, and a simple means of storing words of interest for post-reading investigation. At http://lextutor.ca/hypertext/ ; research and related development program at Cobb (2009).

[B] Writing: Writers in search of the next word or its form for the texts they are writing can click to get a concordance for the last word they managed to get down (or last two, or last three) to see if there are any patterns in the Brown corpus that are applicable to their choice. At http://lextutor.ca/concord_writer/ .

[B] Error feedback: Teachers embed links to selected corpus information beside their learners' collocation or other errors in electronic submissions. Learners correct the error in light of the corpus information. Described and tested in Gaskell and Cobb (2004); at http://lextutor.ca/concordancers/concord_e.html .

[B] Metalinguistic awareness: Learners submit their own texts to Vocabprofile analysis (http://lextutor.ca/vp/) or Keyword analysis (http://lextutor.ca/keywords/) to reflect on their how the lexical properties of their own texts compare to those of other learners or native speakers.

[B] Reaction Time: Learners send their own word selections to a pedagogically adapted lexical decision task instrument as a means to build up their word recognition speed and work toward automaticity. At http://lextutor.ca/rt/ .

[A] Future

It is argued here that spoken language does not normally create stable word representations, and hence that learning vocabulary from spoken language alone takes place slowly and not very extensively. But this picture could change in the near future with the emergence of automatic speech recognition (ASR) technology. If any spoken utterance such as a phone conversation could be simply transformed into an accurate written rendition with a device that was, say, a part of one's mobile telephone, then the vocabulary learning opportunities of speech would be significantly expanded and the distinction between speech and writing significantly reduced. Such a realignment could make a big difference to vocabulary learning, since most people listen to far more language than they read. We are perhaps 10 years away from a highly usable version of such a technology (Levis & Ruslan, this volume).

[A] Conclusion

If the sizes of modern written lexicons are prodigious, the tools available for coming to grips with them are also prodigious. This entry proposes that language technologies and lexicons grow together, - as more words can be learned, more words will need to be learned, in a dynamic growth scenario that shows no sign of abating. But it also proposes what was created by technology can also be managed by it.

Few would dispute that vocabulary is the aspect of language learning that has been most involved with information technology up to now, but they might not agree about the reason. The reason commonly given is that discreet units like words are all that the current state of information technology can handle. A more interesting reason is that technologies and lexicons were intimates long before the advent of computer technology and form a natural pair.

End Notes

1: There are two versions of the argument for the cognitive consequences of writing. Members of the "Toronto school" conclude that the invention and use of phonetic writing caused new kinds of thinking to become possible (the position of Goody, Olson, Havelock Ellis, and McLuhan). Other students of writing effects find in writing mainly the possibilities of supporting, offloading, and extending memory (e.g., Donald, 1991), without any necessity of causality. The argument here is adequately framed by the latter proposal.

ADDITIONAL READING: Automatic speech recognition; Information retrieval for reading tutors; Mobile Assisted Language Learning (MALL); Second Language via Second Life; Text-to-Speech synthesis research; Technology and usage-based teaching applications; Vocabulary load analysis; Web-based lexical resources.

References

Adolphs, S. & Schmitt, N. (2003). Lexical Coverage of Spoken Discourse. Applied Linguistics 24(4), 425-438.

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. Second Language Research 22 (4), 487-497.

Chall, J. & Jacobs, V. (2003). Poor Children's Fourth-Grade Slump. American Educator. http://www.aft.org/newspubs/periodicals/ae/spring2003/hirschsbclassic.cfm

Chung, T. & Nation, P. (2004) Identifying technical vocabulary. System 32 (2), 251-263.

Cobb, T. (In review 1). The challenge of homographs and frequent multi-words. Eurosla Monographs.

Cobb, T. (In review 2). Does Word Coach coach words? For CALICO special on Call Canadian.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. Language Learning & Technology 11(3), 38-63. http://llt.msu.edu/vol11num3/pdf/cobb.pdf

Cobb, T. (2008). What the reading rate research does not show: Response to McQuillan & Krashen. Language Learning & Technology 12(1), 109-114. http://llt.msu.edu/vol12num1/pdf/cobb.pdf .

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 Vol. 2: Reading in Africa: Beyond the School. Kampala: Fountain/African Book Collective. http://www.lextutor.ca/cv/Vol2Sec5Cobb.doc

Coulmas, F. (1989). The writing systems of the world. Cambridge MA: Blackwell.

Cummins, J. (1981). Bilingualism and Minority-Language Children. Toronto: Ontario Institute for Studies in Education. http://eric.ed.gov/PDFS/ED215557.pdf

Dixon, R. (1980). The languages of Australia. London: Cambridge University Press.

Elgort, I. (2007). The role of intentional, decontextualized learning in second language vocabulary acquisition: Evidence from primed lexical decision tasks with advanced bilinguals. PhD dissertation, Victoria University of Wellington, New Zealand.

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

Godwin-Jones, R. (2010). Emerging technologies from memory palaces to spacing algorithms: Approaches to second-language vocabulary learning. Language Learning & Technology, 14(2), 4-11.

Goulden, R., Nation, I.S.P. and Read, J. (1990) How large can a receptive vocabulary be? Applied Linguistics 11, 4: 341-363.

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

Kucera, H. & Francis, W. (1967). Computational Analysis of Present-Day American English Providence: Brown University Press.

Laufer, B. (2005). Instructed second language vocabulary learning: the fault in the 'default hypothesis'. In A. Housen & M. Pierrard (Eds.), Investigations in Instructed Second Language Acquisition (pp. 286-303). Amsterdam: Mouton de Gruyter.

Leech, G, Rayson, P., & Wilson, W. (2001). Word frequencies in written and spoken English: Based on the British National Corpus. London: Longman.

Martinez, R., & Schmitt, N. (In review). A phrasal expressions list.

McQuillan, R. & Krashen, S. (2008). Can free reading take you all the way? A response to Cobb (2007). Language Learning & Technology 12 (1), 104-108. http://llt.msu.edu/vol12num1/pdf/mcquillan.pdf.

Mondria, J.-A. & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of word cards and 'hand computer': Theory and applications. *System, 22,* 47-57.

Nakata, T. (In press). A critical investigation of flashcard software for second language vocabulary learning. Computer Assisted Language Learning.

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

Palmer, H. E. (1917). The scientific study and teaching of languages. London: Harrap. (Reissued in 1968 by Oxford University Press).

Schmitt, N Jiang, X, and Grabe, W. (in press, for 2010). The percentage of words known in a text and reading comprehension. Modern Language Journal.

Segalowitz, N., & Hulstijn, J. (2005). Automaticity in bilingualism and second-language learning. In J. Kroll & A. de Groot (Eds.), Handbook of Bilingualism: Psycholinguistic Approaches (pp. 371-388). Oxford, UK: Oxford University Press.

Tarone, E., Bigelow, M., & Hansen, K. (2007). The impact of alphabetic print literacy level on oral second language acquisition. In N. Faux (Ed.), Low-educated second language and literacy acquisition: Research, policy, and practice (pp. 99-122). Richmond: The Literacy Institute at Virginia Commonwealth University. http://www.theliteracyinstitute.org/projects/pdf/LESLLA%20Research6.pdf

University of Toronto: Dictionary of Old English (DOE) Corpus, 650-1250 CE. http://www.doe.utoronto.ca/about.html .

Webb, S. & Rodgers, M. (2009). The Lexical Coverage of Movies. Applied Linguistics 30(3), 407-427.

West, M. (1926). Learning to read a foreign language (2nd ed.). London: Longmans, Green. (Reissued in 1955 by Longman).