

Center for Advanced Language Proficiency Education and Research
Position paper
October 2007

Assessing Development of Advanced Proficiency Through Learner Corpora

by

Michael McCarthy

CALPER Corpus Project, Director

In the past several decades, the use of language corpora has resulted in considerable advances in our understanding of the quantitative aspects of languages. In the case of English, for instance, most corpora show that around 2,000 word-forms are used very frequently in everyday spoken and written language (O’Keeffe et al, 2007). These word-forms may be said to represent the core of the lexicon. Equally, by measuring how frequent and how widely distributed grammatical structures are in general corpora, we can arrive at a set of core grammatical structures. In the case of the lexicon, the remaining, non-core vocabulary occurs with relatively low frequency but is massive in size (between 30-50,000 word-forms being in use in everyday talk, and considerably more in everyday written texts, perhaps up to 80,000 in the case of English). It would seem non-controversial to suggest that, in language acquisition, the core lexicon and grammar represent a clear target for the elementary level and lower-intermediate level. Language teaching programs have now begun to acknowledge the role of corpora in the definition of these levels; for example, the *Touchstone* adult English language learning series (McCarthy et al, 2004-2006) uses spoken and written corpora to define all four levels of its syllabus from elementary to lower intermediate, as well as evidence from learner corpora.

The advanced level represents a different problem which in the end takes us beyond quantification. In the case of the lexicon, the long tail of non-core vocabulary is simply too big to learn – virtually everything is low frequency, making sub-division and syllabus organisation difficult. It may be possible to say at what point a learner could be said to ‘enter’ the advanced level in terms of number of words (perhaps after 6,000 words have been acquired), but little more can be said in terms of quantity of vocabulary. After the intermediate level, there remains an enormous number of words to be learnt, too many to teach and practise in any secondary or

tertiary language program. The grammar is a more finite entity and 'advanced', in the sense of low-frequency structures, can be delimited and listed in a fairly exhaustive way.

However, what corpora crucially show us is that advanced language proficiency cannot be a simple matter of measuring quantity of vocabulary and grammar known. The evidence of corpora shows that what is at stake in the low-frequency zones is more to do with new kinds of understandings, understandings not only of what is to be learnt, but what one needs to learn about what is to be learnt, and what kinds of learning approaches are most effective for such an open-ended task. The advanced level, therefore, becomes not a teleologic endeavour, but a particular type of development which can be traced over time.

Quantitative features of advanced language proficiency

The purely quantitative aspects of language learning at the advanced level can be represented through frequency lists for both the grammar and the lexicon. For instance, a corpus frequency list can show the increments in text coverage (i.e. comprehensibility using one's receptive vocabulary) offered by adding further 2,000-word bands to the core 2,000. Unfortunately, the leaps required to get to 97% (expert-user level) coverage are not evenly spaced. A 6,000-word upper intermediate vocabulary offers 91% comprehension of typical English texts. Adding another 4,000 words (from the 6,000 to 10,000 word level) accounts for only a 3% gain in coverage, and the next 6,000-word increment (from the 10,000 to 16,000 word level, not shown in the graph) only brings with it a meagre 2% gain, and so on.

Nation (2001:147-8) argues that for full, pleasurable engagement with the meaning of a text, comprehension in the region of 97-98% must be the desired threshold, which would seem to be beyond the grasp of most learners in most language programs, requiring a receptive vocabulary of much greater than 10,000 words. And yet many learners achieve impressive, expert-user levels of knowledge and fluency in all four language skills, either through intensive learning or through periods spent in the target-language environment. Clearly there is more at stake than simply how many words one knows.

On the evidence of corpora, much advanced level vocabulary acquisition will be concerned with less frequent, extended and metaphorical senses of words, and the creation of new relationships among words. The expansion of associations and the forging of new networks is seen as a central aspect of being an advanced learner or user by researchers such as Wolter

(2001; 2002), and Wilks and Meara (2002). Additionally, single-word frequency lists alone do not tell the whole story, and must be weighed alongside the frequency of fixed expressions as evidenced in all corpora. Many fixed expressions equal or exceed in frequency the single-word forms. The ability to retrieve a repertoire of fixed expressions is the hallmark of fluency and must be counted as a central feature of advanced proficiency. Corpora also show that spoken-written differences may be wider and more important at the advanced level. For example, in the case of English syntax, subject ellipsis is of low frequency generally, and especially in written texts (so rendering it outside of the core, high-frequency structures taught at elementary and intermediate levels) but is much more common in informal conversations. Sensitivity to such register- and genre-based distinctions would seem to be a key element of advanced proficiency.

Another characteristic of words in the low frequency bands is that they seem less capable of innocent, neutral use, and a great deal of focus will necessarily be on the connotations of words in their typical contexts of occurrence, over and above grappling with semantic issues. The connotations of words and their characteristic environments of use (their *semantic prosody*, as it has been called; see Sinclair, 1991) seem to be more foregrounded.

Qualitative features: Breadth versus depth

Corpus evidence suggests that the quest for an ever larger and larger vocabulary reflects a rather one-dimensional view of advanced level achievement. A focus simply on linear increase in vocabulary size (or vocabulary *breadth* as it is often termed) produces diminishing returns as far as text coverage is concerned. What needs to happen alongside the increase in breadth is an increase in *depth* of knowledge, i.e. the knowledge of the various aspects of use of a word, including, beyond its formal properties, its collocations, its sub-senses, and its semantic prosody. Such knowledge ultimately contributes to the learner's ability to create associations between words and to place them meaningfully within various networks in relation to other words (Haastrup & Henriksen, 2000; Henriksen, 1999; Meara, 1996).

Depth of knowledge is not simply a second-best to ever-increasing breadth: Qian (2002), for instance, found that vocabulary depth was as significant as vocabulary size in predicting performance on academic reading. And since the vocabulary learning task is open-ended and impossible to complete in a typical institutional program, the implication is that the advanced level should also be defined by the extent to which the learner is able to operate independently

with a set of skills and strategies for processing new vocabulary at this level. Such a learner may not have a massive vocabulary but may be better equipped to use and explore the vocabulary of the target language than one who simply adds more and more words without building an integrated lexicon and without developing that ‘learner agency’ so often discussed in sociocultural theory (Lantolf and Appel, 1994), which can enable the learner to surpass instructional intervention and become a better, self-regulated learner.

Conclusion

To develop awareness and skills that will stand the learner in good stead for becoming an autonomous vocabulary-learner is a question of developing activities alongside the actual learning of words which introduce to the learner notions such as collocation, metaphor, connotation, etc. For example, in the case of English, many learners have an awareness of idioms of the *verb+complement* type (*hit the sack, carry the can, jump on the bandwagon*), but probably few are aware of the pervasiveness in everyday language of binomial idioms (*rough and ready, part and parcel, out and about, down and out*). Explicit focus on such items may be necessary to tune the learner’s antennae to be receptive to new ones, and to foster learner agency and independence. Vocabulary skills include ways of maximising learning opportunities during interaction (e.g. asking for paraphrases, probing the meaning of unfamiliar items with one’s interlocutor, etc.).

The advanced level learner will not be defined only by his/her vocabulary size or absolute coverage of all syntactic patterns vis-à-vis native speakers, but rather more by his/her ability to develop depth of knowledge and the tools and strategies to pursue vocabulary learning independently. With a combination of corpus-based research and the pursuit of strategic training for learners who will have to complete the task for themselves, we may go a long way towards defining advanced language proficiency in terms of traceable development on axes other than the purely quantitative increase of acquired structures and words.

References

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied*

- Linguistics* 10(2): 221-240.
- Henriksen, B. (1999) Three dimensions of vocabulary development. *Studies in Second Language Acquisition* 21: 303-317.
- Lantolf, J. P. and Appel, G. (eds.) (1994) *Vygotskian approaches to second language research*. Norwood, NJ: Ablex.
- McCarthy, M. J., McCarten, J. and Sandiford, H. (2005-2006) *Touchstone*. Student's Books 1-4. Cambridge: Cambridge University Press.
- Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press.
- Nation, I. S. P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- O'Keeffe, A., McCarthy, M. J. & Carter, R. A. (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.
- Qian, D. D. (2002) Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52 (3): 513-536.
- Wilks C. and Meara P. (2002) Untangling word webs: graph theory and the notion of density in second language word association networks. *Second Language Research* 18 (4): 303-324.
- Wolter, B. (2001) Comparing the L1 and L2 mental lexicon: a depth of individual word knowledge model. *Studies in Second Language Acquisition* 23(1): 41-69.
- Wolter, B. (2002) Assessing proficiency through word associations: is there still hope? *System* 30(3): 315-329.

The Center for Advanced Language Proficiency Education and Research (CALPER) at the Pennsylvania State University is one of 15 National Language Resource Centers funded by the U.S. Department of Education (CFDA 84.229 P229A060003).