# Collocations in a Learner Corpus (review)

Tom Cobb



The Canadian Modern Language Review / La revue canadienne des langues vivantes, Volume 63, Number 2, December / decémbre 2006, pp. 293-295 (Review)

Published by University of Toronto Press
*DOI:* *https://doi.org/10.1353/cml.2007.0001*

# Book and Software Reviews / Critiques de livres et de logiciels

**Nesselhauf, Nadja.** (2005). *Collocations in a Learner Corpus.* Amsterdam: John Benjamins. Pp. xii, 332. US$126.00 / €105.00 (cloth).

If vocabulary was the flavour of the month in applied linguistics in the last decade, in this one it is the multiword unit (MWU). This refinement follows a certain logic, because, if the claims made for said unit are even half accurate, then all levels of the language teaching industry are in for a significant rethink. Ideas to be incorporated would be that grammars emerge from phrases, not vice versa; that main tasks in language acquisition are piecemeal, not rule based; and that functioning lexicons consist not in manageable handfuls of words but in vast array of combinations lexicalized to varying degrees and operating within mazes of apparently random restrictions. No surprise that the rethink has hardly begun, with progress somewhat held up, until recently, by the lack of clear terms and an empirical database. To contribute to ongoing work on both fronts is the purpose of Nadja Nesselhauf's book, which is based on her doctoral study of one type of multiword unit in the written production of advanced German-speaking ESL students.

Nesselhauf has assembled a corpus of advanced learner writing with a view to inspecting one of its MWUs, following procedures for learner corpus research established by Granger (1998). But Nesselhauf goes beyond anything published to date in her delimitation of phenomena and her generation of comparable data. In a detailed but (largely) readable account of her methodology, she carefully separates out collocation as the type of MWU she will look for, and, within that category, verb–noun collocations (i.e., *ride a bike* not **drive a bike*), with the specification that the restriction (on *ride*) be fully arbitrary rather than meaningful. A catalogue of such collocations is hand-extracted from her learner corpus, and native collocations separated from learner deviations by native raters; collocations are counted in terms of

© 2006 *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 63,2 (December/décembre), 293–306

frequency and range, and deviations are categorized by type and probable intended meaning. All data can be traced back through individual writers to a background questionnaire itemizing years of ESL study, extent of exposure to English abroad, and conditions of writing (e.g., timed or untimed; dictionary yes or no). To call this 'a lot of work' is an understatement, and, indeed, the amount of manual work involved raises the question of whether this approach can be scaled up to a larger corpus (than the 200,000 words used here), as Nesselhauf proposes.

But even a smallish corpus, carved with instruments so fine, can generate interesting information. A predictable finding is that collocation remains a serious problem well into advanced learning. Less predictable is that neither years of instruction, nor years abroad, nor writing with or without time pressure, with or without a dictionary, has any effect on the number of collocations employed or on the number of deviations. Particularly interesting are the deviations exposed by imputing intended meaning – as when a learner writes 'I don't take care of carrots,' which is a good collocation, except that he probably means 'I don't care for carrots' (which a computer match of learner strings against a standard corpus would have missed).

So, then, the problem is even worse than we thought; but what is the solution? Nesselhauf explores awareness versus learning as solutions. Learners are not (encouraged) in the habit of scanning language to become aware of restrictions on word combination. The collocation problem resembles one from the vocabulary research: that a word met in rich contexts can have a meaning so obvious that the word itself does not register in memory – 'ride a bike' paints a picture so clear that there is little motivation for the learner to notice that it was *ride* and not *drive.* A lengthy section on pedagogical implications suggests ways of promoting awareness as well of developing a collocational syllabus.

Interspersed in Nesselhauf's treatment are attempts to clarify unresolved issues in the MWU agenda. One concerns Kjellmer's (1991) idea that, while natives process language in prefabricated sequences, learners rely on grammars and lexicons, which leaves them 'sounding odd.' A problem is that if fluency is impossible without access to MWUs (Sinclair, 1991), but learners do become fluent users of second languages, then either they employ such units or else fluency can in fact be achieved on a words-and-rules basis. One way through the paradox is the frequent finding (e.g., Cobb, 2003) that learners do use MWUs, including collocations, and, indeed, overuse the few that they have, which is why their language sounds odd. Nesselhauf proposes another angle on the Kjellmer question, which there is no space to mention here (and, anyway, a review should not give away too much).

It is to be hoped that the clear thinking and methodological exactitude of Nesselhauf's study will be taken up in further studies. Should others accept the challenge to advance the MWU agenda through hard and careful work, as Nesselhauf has done, there are some things to watch for in writing up the results. First, while detail and precision clearly advance the research, the reading can be heavy going (e.g., pp. 240–241 offer two pages of closely reasoned linguistics with just one example to give a breather). Second, some of the apparatus of a thesis is out of place for a book audience (e.g., 30 pages of endnotes). Third, with all the pains taken to quantify her data, Nesselhauf nonetheless relies entirely on descriptive statistics, even when making comparisons (e.g., between collocation counts in timed and non-timed writing). She even describes the results of comparisons with folkloric expressions. A longer stay in an English-speaking country 'does not seem to lead to' an increased use of collocations (p. 236); the percentage of deviant collocations for users and non-users of dictionaries was 'exactly the same' at 36.1% (p. 231). Isn't the point of *t*-tests to tell us which differences really are different?

These criticisms simply mean that no study can do everything and that much work remains to be done in this area. The methods developed here are eminently replicable; the holes to plug are obvious. This is not easy research, but careers will be made in it – this is the first act in a drama that will unfold for years to come.

### References

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *The Canadian Modern Language Review, 59,* 393–423.

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

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

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

**Tom Cobb**, *Université du Québec à Montréal*

**Price, Glanville.** (2005). *An Introduction to French Pronunciation.* Malden, MA : Blackwell. Pp. 176, 29.95 $US.

Le livre de Glanville Price intitulé, *An Introduction to French Pronunciation*, s'adresse à ceux qui ont déjà une certaine connaissance de la