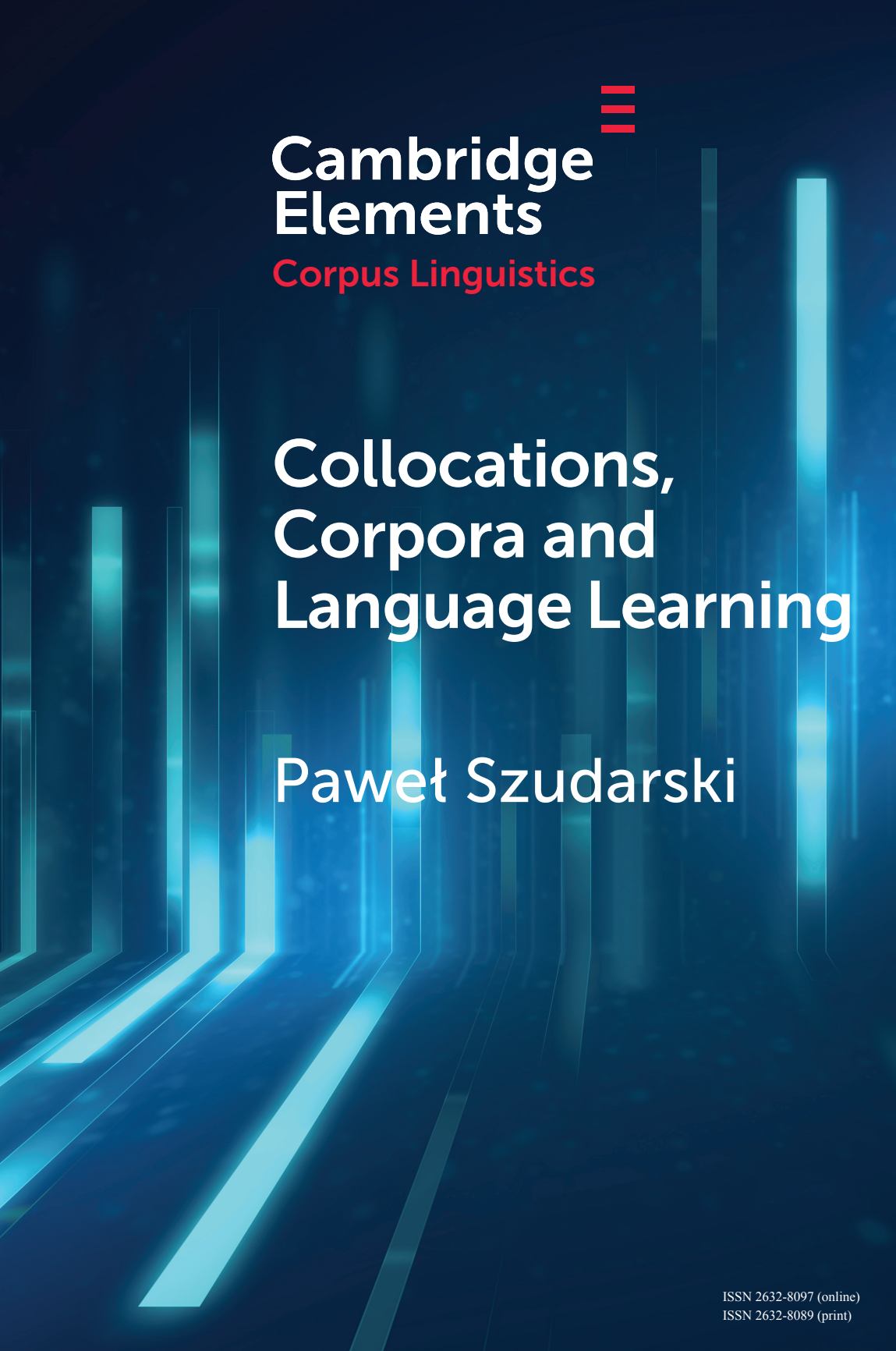


ISSN 2632-8097 (online)
ISSN 2632-8089 (print)

Cambridge Elements

**Elements in Corpus Linguistics**
edited by
Susan Hunston
*University of Birmingham*

# COLLOCATIONS, CORPORA AND LANGUAGE LEARNING

Paweł Szudarski
*University of Nottingham*

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108994798

DOI: 10.1017/9781108992602



First published 2023

*A catalogue record for this publication is available from the British Library.*

ISBN 978-1-108-99479-8 Paperback
ISSN 2632-8097 (online)
ISSN 2632-8089 (print)

# Collocations, Corpora and Language Learning

Elements in Corpus Linguistics

DOI: 10.1017/9781108992602
First published online: June 2023

Paweł Szudarski
*University of Nottingham*

**Author for correspondence:** Paweł Szudarski, pawel.szudarski@nottingham.ac.uk

**Abstract:** This Element provides a systematic overview and synthesis of corpus-based research into collocations focusing on the learning and use of collocations by second language (L2) users. Underlining the importance of collocation as a key notion within the field of corpus linguistics, the text offers a state-of-the-art account of the main findings related to the applications of corpora and corpus-based measures for defining, identifying and analysing collocations as related to second language acquisition. Emphasising the quality of L2 collocation research, the Element illustrates key methodological issues to be considered when conducting this type of corpus analysis. It also discusses examples of pertinent research questions and points to representative studies treated as models of good practice. Aiming at researchers both new and experienced, the Element also points to avenues for future work and shows the relevance of corpus-based analysis for improving the process of learning and teaching of L2 collocations.

**Keywords:** collocation, corpus analysis of collocations, learning collocations in a second language, corpora and phraseology, formulaic language



ISBNs: 9781108994798 (PB), 9781108992602 (OC)
ISSNs: 2632-8097 (online), 2632-8089 (print)

# Contents

## 1 Introduction

The last few decades have seen an impressive growth of interest in corpus-based analysis of language. Corpora, large computerised collections of language data, have been instrumental in the expansion of many subdisciplines within linguistics, and it is fair to say that corpus methods have become an indispensable tool for much of contemporary linguistic research. In fact, in appraising the impact of corpora on the way linguistic analysis is currently carried out, some authors (e,, Hanks, 2012; Chambers, 2019) have gone as far as stating that corpora have revolutionised language studies by providing a whole new range of methods and tools to study language, its learning and use across varied settings and contexts (see also O'Keeffe & McCarthy, 2022, for a summary of the evolution of corpus linguistics in the last ten years).

Examples of areas where corpora have proved highly useful are plentiful and include, among others, corpus-assisted discourse analysis, register and genre variation studies, second language acquisition (SLA) and applied corpus-based research. This Element focuses on the last two, underlining the importance of corpora for exploring collocations as a type or category of the broader phenomenon of formulaic language (see Section 2.1 for an overview). Specifically, the discussion centres on corpus-based and corpus-informed analyses of collocations treated as frequently recurring two-to-three-word lexical units characterised by relative transparency of meaning and restricted connectedness of the constituent words (e.g., 'make an error' as opposed to 'do an error'). In particular, the Element demonstrates the pivotal role of corpora in analysing collocations as related to second language (L2) research, offering a critical synthesis of the current findings and pointing to key methodological considerations that affect the quality and validity of collocation studies.

My intention as the author of this text has been to provide a useful account of the main concepts and debates in the field, not only by presenting the most pertinent research questions and examples of studies in this line of inquiry but also by discussing the key methodological decisions that need to be made in carrying out this type of corpus-based work. By overviewing the main traditions and approaches followed in collocation studies, the Element seeks to present specific methods and types of analysis, explaining how corpus data, methods and tools are particularly effective at delving into the varied ways in which words co-occur and collocate as phraseological partnerships. In this sense then, *Corpora, Collocation and Language Learning* has been conceptualised as a specialised but accessible introduction to corpus-based collocation research, aimed at both fellow linguists interested in studying the phenomenon of collocations but also language practitioners who may want to turn to corpora as a way

of addressing practical challenges linked to selecting and teaching examples of specific word pairs deemed important for L2 pedagogy.

In practical terms, this means that by the end of the Element, the reader should have a thorough theoretical understanding of collocations as a key concept in corpus linguistics. They should also be well versed in the mechanics and methodologies associated with corpus-based analyses of collocations, enabling the pursuit of questions like the following:

- How do we define and identify different types of collocations?
- Which collocations are used more often in learner language or academic language?
- What is the relationship between the frequency of occurrence and the learning of collocations? What other factors affect this process?
- How is the learning and use of collocations by advanced L2 learners different from that by intermediate-level learners?
- What is the relationship between the use of collocations and the assessment of L2 learners' proficiency?
- How can corpus research inform the process of L2 teaching and materials development so that learners are provided with the optimal conditions for learning collocations?
- What aspects of L2 learning and teaching can benefit from the affordances of corpus analysis?

With these questions serving as the starting point, the Element is divided into five sections. Following this Introduction, Section 2 focuses on defining the term 'collocation', situating it in the literature on the broader phenomenon of formulaic language and explaining how corpora have been pivotal in advancing the understanding of this topic. Section 3 is an overview of the main corpus methods and tools that can be applied to the study of collocations. By discussing aspects of corpus analysis and presenting representative corpus-based studies, the aim of this section is to show how to search for examples of collocations in corpora, apply different corpus-based measures and statistical tests of word partnerships and analyse the use of collocations by taking multiple research perspectives. Building on this, Section 4 lies at the heart of this Element and provides a selection of corpus-based studies into L2 collocational learning and teaching, focusing specifically on learner corpora and a variety of factors that affect the acquisition and use of collocations by first and second language (L1 and L2, respectively) speakers. Linking corpus insights with findings from SLA, psycholinguistics and language pedagogy, this section showcases key findings in L2 collocation research, presents exemplary studies which model how to draw on corpora and discusses the practical implications of this work for

language education. Finally, Section 5 offers a summary of the Element and recognises contributions and recent developments within corpus-based analysis in terms of advancing collocation studies and applied linguistics research more broadly. Using the reviewed findings, the discussion concludes with reflections on the evolution of the field, emphasises the instrumental role of corpora in studying collocations as a crucial aspect of language and describes possible avenues for future empirical work.

To aid the reading process, the Element includes a number of features whose aim is to not only enhance the reader's understanding of the main issues but also to invite them to critically engage with the existing collocation studies and consider the numerous methodological choices that need to be made as corpus analysis is undertaken. One such feature is quotations presented throughout the text, which illustrate the main points being discussed; another is study boxes which report relevant findings from corpus-based studies and model best practice in carrying out collocation analysis. Further, considering the current popularity of corpus-based work into the collocability of words and L2 phraseology, the Element also references many examples of studies which can be consulted for further information, with a view to encouraging readers to engage with the wider literature, immerse in the richness of corpus-based inquiry and embark on their own journey in this fast-growing area of linguistic analysis.

Finally, it is also worth adding that while the Element centres on collocations (e.g., 'make a mistake', 'strong coffee', 'extenuating circumstances'), where relevant, the discussion also draws on the wider spectrum of corpus-based work into formulaic sequences broadly defined as 'multiword phenomena which holistically represent a single meaning or function' (Wood, 2020, p. 30). In such cases, it is clearly indicated which types of phrases or formulaic units are being referred to, with explanations provided on how specific types of corpus-based analyses contribute to broadening our understanding of word co-occurrence, formulaicity and phraseological patterning.

## 2 How Are Collocations Defined?

### 2.1 Collocations and Formulaic Language

In the last thirty years or so, there has been a great deal of attention paid to collocations as a key element of language, with important developments in corpus analysis resulting in a multitude of new research focused on vocabulary studies (Szudarski, 2018; Granger, 2021; Durrant et al., 2022; Szudarski & Barclay, 2022). Thanks to the advent of corpora, it has become clear that language is highly patterned and to a large extent consists of fixed vocabulary and phraseological units, including not only collocations but also idioms ('red herring'),

binomials ('ladies and gentlemen'), lexical bundles as contiguous sequences of words that recur in speech and writing ('it is important that') and other types of phrases (for a useful discussion of research into such multi-word units, see Siyanova-Chanturia & Omidian, 2020). In fact, the discovery that multi-word units are ubiquitous in natural language has been one of the major contributions of corpora to the field (Forsberg Lundell, 2021), bringing a new vitality to lexical studies and resulting in 'a complete overhaul of the theory and practice of phraseology' (Granger, 2021, p. 5).

With this in mind, this section focuses on questions related to defining collocations, recognising the importance of corpora in identifying relations between collocating words and explaining also how collocation studies need to be considered within the broader context of corpus-based research. That said, while this Element is very much grounded in the wider discussion devoted to the formulaicity of language, it is important to note that its goal is not to provide a detailed review of the vast literature devoted to this topic (for a comprehensive account, see Siyanova-Chanturia & Pellicer-Sanchez, 2019; see also Schmitt, 2022 for a useful summary). Rather, after this introductory section, the remainder of the text is concerned predominantly with collocations treated as a type of formulaic language, with examples of specific pairs of words identified according to both phraseology- and corpus-based criteria (for details, see Section 2.4).

In terms of the structure of this section, the paragraphs that follow first present collocation as a central concept in corpus linguistics, with Section 2.2 relating collocation research to Sinclair's idiom principle and the terminological challenges besetting this area of work. Next, the importance of corpora is underscored, highlighting their role in studying the graded and probabilistic nature of collocations as observed in the lexical and lexico-grammatical partnerships they form (Section 2.3). Crucially, whilst individual language users can identify such partnerships in informal and intuitive ways, it is also true that their subjective intuitions and predictions might turn out to be inaccurate or inconsistent. For instance, when it comes to rating lower-frequency words and phrases, research points to variation in the consistency and accuracy of responses amongst both L1 and L2 speakers (Schmitt & Dunham, 1999; Alderson, 2007; Siyanova-Chanturia & Spina, 2015). This is where the power of corpora comes to the fore, with Section 2.4 describing the main traditions followed in collocation research and introducing a range of measures used in corpus-based studies. Not only do they allow us to measure collocations in a reliable and automatic way but they also help to tap into different dimensions of word co-occurrence, throwing light on the intricate ways and relations between collocating words. Yet another dimension of collocation studies is tackled in Section 2.5, which makes a distinction between the textual and the psycholinguistic reality of collocations.