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.

2.2 Collocations as a Central Concept, Idiom Principle and Terminological Challenges

A great deal of corpus-based research into collocations has been inspired by the pioneering work of John Sinclair, one of the key figures in establishing corpus linguistics as a new research paradigm and method of analysis. Based on the 1980s work on the COBUILD Corpus, Sinclair (1991) proposed that language and its use are governed by two main principles: the open-choice principle, in which speakers are unrestricted in their linguistic choices and construct sentences by selecting words item by item; and the idiom principle, according to which speakers and writers construct sentences and utterances by means of 'ready-made' phrases, collocations and phraseological chunks. This Element focuses on the latter, putting collocations at the centre of language analysis and highlighting their role in L2 learning and use.

Quote 1
Firth (1957, p. 179) 'You shall know a word by the company it keeps.'

When it comes to research, Sinclair and his followers have been less preoccupied with setting clear boundaries and determining whether phraseology is concerned more with lexis or grammar. Rather, they have emphasised the pervasiveness of collocations and other types of word combinations, stressing the part such phraseological units play in conveying specific meanings and fulfilling important pragmatic functions. This, for instance, includes phrases that express speakers' stance or attitudinal meanings (for details on discourse functions performed by collocations and phraseological units, see Schmitt & Carter, 2004; O'Keeffe et al., 2007; Carter, 2012; see also Durrant & Mathews-Aydinli, 2011 for details on a function-first approach to the identification of important phrases). Put differently, corpus-based accounts of phraseology have emphasised the idiom principle and the concept of lexico-grammar (O'Keeffe et al., 2007; Römer, 2009; Szudarski, 2018). In this approach, rather than seen as two separate levels of language, grammar and lexis are intrinsically intertwined (Granger, 2021), with recurrent collocations and lexico-grammatical patterns occupying a central position and reflecting a distinct psycholinguistic reality of phrases in speakers' mental lexicons (see Section 2.5 for details on Hoey's (2005) notion of lexical priming and the psychological status of collocations).

Importantly, while useful in terms of framing the discussion, wide-ranging notions such as 'formulaic language' or 'idiom principle' can also be problematic (Myles & Cordier, 2017). One important issue is the multiplicity of terms that have been employed in the literature on phraseology. For instance, in her seminal publication devoted to this topic, Wray (2002) identified as many as

forty different terms that had been used in studies devoted to the formulaicity of language. Granger (2009) has rightly referred to this as a terminological chaos which besets this strand of research.

A related challenge is that the literature abounds in a wide range of definitions that encompass different types of phrases, without necessarily doing justice to how items presented as collocations across specific studies might in fact differ from each other along the key dimensions of formulaicity such as fixedness, non-compositionality, familiarity or L1–L2 congruence. A case in point is an oft-cited study by Webb et al. (2013), which examined the process of incidental learning of L2 collocations. The target items in this study included phrases of varied phraseological status (e.g., 'cut corners', 'pull strings', 'throw light'), which, in the light of the non-literal meanings of some of these phrases, could arguably be categorised as idioms rather than collocations. Such examples then show how the categorisations of specific phrases are highly dependent on individual scholars' decisions to adopt specific criteria and definitions (see Peters et al., under review, as an instance of a study which deliberately uses the term 'multiword units' as a broader category).

Commenting on such difficulties in defining and categorising examples of formulaic sequences, Wood (2020) observes that it is common for authors to hedge their claims about formulaic language and the multifaceted nature of research in this area. Similarly, Granger (2021) points out how different operationalisations and definitions found in phraseological research render it difficult to compare findings across studies, particularly if data collection involves different designs or research protocols.

Taking all of the above into account, this Element treats collocations as a category of formulaic language but underlines the importance of establishing clear and replicable definitions, with corpora treated as a useful source of insights into the graded nature of collocability. The next section specifically explains how the identification of collocations can benefit from corpus-derived information.

2.3 Defining Collocations as a Graded Phenomenon

Having introduced the wider context for this Element, it is important to say that collocation is a topic that has been attracting increasing amounts of attention in corpus linguistics and beyond. Gries (2013, p. 159) goes as far as stating that 'collocation has been, and will remain, one of the most important concepts' in corpus-based inquiry. At the same time, he also calls for careful consideration of how collocations can be defined depending on specific research purposes. Similarly, Gyllstad and Wolter (2016) observe that when one looks at published

collocation research, what is referred to as a collocation varies greatly both within and across studies, which compounds the task of adopting appropriate and commonly accepted definitions.

Broadly speaking, collocations can be regarded as word partnerships, with collocation research putting a strong emphasis on different types of lexical and lexico-grammatical relations and patterns. Specifically, corpus-assisted analyses of the lexical company of individual items help to reveal new facts not only about word co-occurrence, but also recurrence, metaphoricity, creative use of language and many more. Also, with linguistic knowledge and use viewed as a formulaic-creative continuum (Ellis et al., 2015), analyses of the collocability of words help to investigate and identify collocations as a graded phenomenon with different degrees of probability (Sinclair, 2004).

Quote 2
Sinclair et al. (2004, p. 72) 'There is no hard and fast distinction between a casual and regular collocation, simply different degrees of probability.'

For instance, such a graded view of collocations can be found in Granger's (2021) discussion of learner corpus research (for details on learner corpora, see Section 4.2). Rather than employing binary categories of collocations versus non-collocations, the author calls for approaching L2 learner production by means of a range or cline, where pairs of words may be more or less strongly associated and psycholinguistically entrenched. From this perspective then, if collocations are a graded phenomenon, the key task for individual corpus users working across different data sets is to ensure that clear and reproducible criteria are applied, leading to more consistent and comparable findings (e.g., comparisons of collocations retrieved from corpora that represent different L2 learning and teaching contexts).

This is where Gries's (2008; 2013) work is highly relevant, because he offers a number of recommendations to be taken into account while conducting research into collocations and other phraseologisms. He approaches phraseologisms in a rather broad way, defining them as 'the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which function as a semantic unit in a clause or sentence and whose frequency of occurrence is larger than expected on the basis of chance' (Gries, 2008, p. 6). Such a broad perspective is inevitable when the ambition is to encompass as many types of phraseological units as possible. Additionally, Gries also argues that we need to be mindful of several dimensions of word co-occurrence phenomena if the field is to ensure comprehensibility, comparability and replicability of findings across different studies:

The nature of the elements involved in a phraseologism
The number of elements involved in a phraseologism
The number of times an expression must be observed
The permissible distance between the elements involved
The degree of lexical and syntactic flexibility of the elements involved
The roles semantic unity and semantic non-compositionality/non-reductability play in the definition

As Gries (2013, p. 136) stresses, these considerations underlie 'most of the work using collocations' and therefore constitute a good starting point to adopt more rigorous definitions and improve the quality of corpus-based phraseological research. By way of example, for a corpus linguist searching collocations in a given corpus, there are a number of questions they are likely to face:

- How many words should I include or allow in my definition of a collocation? For most linguists, collocations are typically two content words as in 'powerful car', but for phrases such as 'make a mistake', it is also important to decide how to treat articles or other grammatical elements.
- Which words classes does my definition of collocations focus on (e.g., verb–noun collocations vs. adjective–noun collocations)?
- How many times should a given collocation occur in my corpus before it can be included as a target item worth studying by an L2 learner? Put differently, what is the required minimum frequency threshold (e.g., 10 times per million words)?
- If individual elements of a collocation occur in different forms (e.g., the collocation 'make a mistake' realised as the forms 'make', 'made', 'making'), do I include all of them under the lemma [make] as my unit of analysis or do I search for each of these forms separately?
- Does my search include only contiguous (adjacent) elements or should my collocation window allow some empty slots to take into account syntactic flexibility (e.g., passive voice constructions in collocations such as 'mistakes are made')?

While highly relevant for corpus-based explorations of L2 phraseology, regrettably most of these questions are not easy to answer, which constitutes a challenge from the research point of view. In fact, the idea of writing this Element was partly motivated by such recurring challenges related to, for instance, establishing replicable criteria in defining and identifying examples of specific collocations. What follows then is a summary of the current debate and main approaches adopted in collocation research, reflecting upon and synthesising the ever-increasing number of empirical studies investigating the

notion of collocations (for a collection of collocation studies, see Barfield & Gyllstad, 2009; for a short overview specifically focused on L2 collocations, see Szudarski, 2017).

2.4 Main Traditions in Collocation Research

2.4.1 Phraseology-Based Definitions

As aptly summarised by Gyllstad and Wolter (2016, p. 297), 'different definitions of what a collocation is abound in the literature', but it is possible to discern two dominant trends or research strands: a phraseological approach and a frequency-based one. The phraseological approach is considered a more traditional school of thought; it is typically associated with Eastern European research into phraseology and employs linguistic criteria as the basis for the categorisation of phraseological units (e.g., Cowie, 1994). Specifically, this approach relies on features such as semantic transparency, restrictedness and phraseological specialism to assess the collocability, idiomaticity and connectedness of specific phrases. As Howarth (1998, p. 27) emphatically states, phraseological significance, due to its complexity, is something less tangible than any computer algorithm can reveal, implying that intuition and judgement on the part of the analyst are necessary to study relations between co-occurring words and discern the numerous ways in which they can be combined. Inevitably, since semantic transparency or restrictedness are not clear-cut criteria, such judgement of phraseological significance involves a large degree of subjectivity, giving rise to different interpretations and consequently affecting the applicability or consistency of this approach (Granger, 2021).

Howarth (1996; 1998), for instance, is one of the authors often cited as representative of this line of phraseological research. In his continuum model of phraseology, he lists the following four categories of word combinations: free combinations ('blow a trumpet'), restricted collocations ('blow a fuse'), figurative idioms ('blow your own trumpet') and pure idioms ('blow the gaff'). Thus, with respect to the fixedness of phrases for instance, this model operationalises collocations as arbitrarily restricted pairs of words, which are more fixed than free combinations but less fixed than idioms. In turn, as regards meaning decoding, collocations are seen as more transparent than non-compositional and often metaphoric idioms.

Importantly, while not devoid of subjectivity in terms of delineating between the specific types of phrases, this continuum model has been influential in inspiring collocation research, mostly because of its great descriptive value and the potential to make linguistic distinctions. A good illustration of this is Gyllstad and Wolter's (2016) study, which used Howarth's model to test the way

different types of word combinations are processed by L1 and L2 users. Study Box 1 presents details of this study.

Study Box 1

Gyllstad and Wolter (2016)

Background & Aims: Considering numerous ways in which words can combine with each other as phrases, there are different linguistic criteria (e.g., semantic transparency or levels of fixedness) that are used to distinguish between free combinations ('pay a bill'), collocations ('pay a visit') and idioms ('pay the piper'). The study sought to investigate the processing of such different combinations in the L1 and L2, as indicated by participants' responses to a semantic judgement task. From the perspective of corpus linguistics, it is worth stressing that the linguistic properties of all the target items were controlled for, including phrasal frequency in the Corpus of Contemporary American English (COCA), length of phrases, the number of their cognates in participants' respective languages and collocational congruency understood as L1–L2 translation equivalence of the constituent co-occurring words.

Research Question(s)

1. For advanced L2 users of English, is there a processing cost for collocations compared to free combinations in terms of reaction times and error rate values?
2. Is the pattern the same or different for L1 users?
3. Is Howarth's descriptive distinction between free combinations and collocations in the continuum model reflected in processing differences?

Methodology

- Twenty-seven L1-Swedish advanced-level users of English and thirty-eight L1-English users
- Reaction times and error rates measured in response to three types of phrases: free combinations ('kick a ball'), collocations ('draw a conclusion') and baseline items (combinations of random words)
- All target items were congruent between Swedish and English to avoid cross-linguistic influence (for a detailed discussion of congruency effects, see Wolter & Gyllstad, 2013)

Results & Discussion: For both groups of participants, processing collocations was found to be more demanding than processing free combinations, revealing that the phraseological features and status of phrases