Article

Stefania Spina*, Irene Fioravanti, Fabio Zanda, Luciana Forti, Damiano Perri and Osvaldo Gervasi

Developing a learner dictionary of collocations: description and evaluation of a multi-method approach

https://doi.org/10.1515/cllt-2025-0008
Received January 17, 2025; accepted January 1, 2026; published online January 13, 2026

Abstract: This study describes and evaluates a multi-method approach for identifying and extracting collocations to develop a learner Italian collocation dictionary. The approach integrates part-of-speech tagging and dependency parsing to extract six syntactic relations from a reference corpus of Italian. The initial set of candidates was gradually reduced using frequency, dispersion, and association measures. This set was then evaluated by comparing it with existing collocation dictionaries and gathering expert judgments on which collocations should be included. Combining these two evaluations, further refined the list. Moreover, the effect of statistical measures on expert judgments was investigated. Results revealed that dispersion and association measures positively influenced human evaluations, while higher frequency often correlated with negative ratings. This triangulation of corpus-based and statistical methods, human judgements and comparison with existing dictionaries captures collocations widely used across genres, suitable for inclusion in a learner dictionary, offering a useful tool for learners while contributing to corpus-based collocation research.

Keywords: collocation; learner dictionary; L2 Italian; frequency; dispersion; association measures

***Corresponding author: Stefania Spina**, University for Foreigners of Perugia, Piazza Fortebraccio, 4, 06123, Perugia, Italy, E-mail: stefania.spina@unistrapg.it. https://orcid.org/0000-0002-9957-3903
Irene Fioravanti, Fabio Zanda and Luciana Forti, University for Foreigners of Perugia, Piazza Fortebraccio, 4, 06123, Perugia, Italy. https://orcid.org/0000-0001-5182-9394 (I. Fioravanti). https://orcid.org/0000-0001-5183-9613 (F. Zanda). https://orcid.org/0000-0001-5520-7795 (L. Forti)
Damiano Perri and Osvaldo Gervasi, Department of Math and Computer Science, University of Perugia, Via Luigi Vanvitelli, 1, 06123, Perugia, Italy. https://orcid.org/0000-0001-6815-6659 (D. Perri). https://orcid.org/0000-0003-4327-520X (O. Gervasi)

1 Introduction

In this study we describe and evaluate an approach for the identification and extraction of Italian collocations with potential dictionary relevance, using a reference corpus (i.e., the PEK24 corpus). The context is the compilation of a Learner Dictionary of Italian Collocations (*Dizionario delle Collocazioni Italiane per Apprendenti*: DICI-A), which is aimed at creating a new lexical resource in the area of Italian lexicography, where none of the three existing collocation dictionaries (Lo Cascio 2012, 2013; Tiberii 2018 [2012]; Urzì 2009) strictly adopts corpus-based methods, nor are any of them specifically targeted at L2 learners of Italian. Creating a lexical resource targeted at L2 learners means contributing to the strengthening of Italian L2 learning. The dictionary will be a suitable tool for individual learning as well as for classroom teaching activities. Indeed, developing a corpus-based dictionary for learners means providing the resource with authentic data extracted from real contexts of Italian usage, which allows the dictionary to be a valuable pedagogical tool to support teachers in carrying out their teaching activities.

Although a vast amount of research over the past two decades has addressed this issue (e.g., Deng and Liu 2022; Evert et al. 2017; Pecina 2010; Seretan 2011; Wahl and Gries 2018), the identification and extraction of collocations in corpora is still an area with much debate about the methods to be adopted. This lack of agreement suggests the adoption of hybrid strategies, integrating the potential advantages of different methods.

Our approach, therefore, relies on the following theoretical assumptions: (i) collocations are a phenomenon at the interface between lexis and grammar (Römer-Barron and Schulze 2009); collocations are thus lexical combinations with a grammatical configuration and syntactic relations that their components preserve even when they are not adjacent; as a consequence, for their detection in corpora it is reasonable to use methods integrating both the ability to identify grammatical sequences and to detect syntactic relations between lexical elements, even at a distance from each other (Castagnoli et al. 2016; Perri et al. 2024); (ii) collocations are domain- and register-dependent (Gablasova et al. 2017); high-dispersion collocations are acquired earlier than domain-specific ones (Chen 2015), and both frequency and dispersion affect their productive use by learners (Candarli 2021), as well as strength of association (e.g., Durrant and Schmitt 2009; Gablasova and Brezina 2025); as a consequence, any extraction method should be based on corpora representing as many different written and spoken genres as possible, and, after their extraction from corpora, the filtering and ranking of collocations should rely on integrated criteria, combining measures of how frequent and dispersed they are in corpora and how strong and exclusive their mutual relation is (Ballance 2022; Gries 2024).

Given these assumptions, we adopt a frequentist approach to collocations (Evert 2009) and use a corpus-based three-step identification method with the following key features: Step A – we integrate part-of-speech (henceforth, *pos*) tagging and parsing techniques for the extraction, from a corpus, of candidate collocations, which in our view are continuous or discontinuous pairs of pos sequences, bound by a syntactic relation. In order to account for register variation, we perform their extraction in a written and spoken reference corpus of Italian including 10 different textual genres (Spina et al. 2025); Step B – we refine frequency counts with bag-of-words techniques; Step C – we integrate measures of frequency, dispersion and statistical association with the aim of retaining collocations with statistical salience which are commonly used in a range of textual genres. Finally, we further filter the candidate collocation list resulting from the corpus-based component of our method, by comparing the items in the list with the entries contained in existing collocation dictionaries and also by using human ratings to evaluate their suitability for inclusion into a learner dictionary. As our candidate collocations were identified using frequency, dispersion and association measures (henceforth, AMs), we investigated to what extent human ratings are predicted by corpus-based metrics.

The present study is structured as follows: Section 2 reviews previous research on the identification and extraction of collocations in language corpora, specifically focusing on lexicographical applications; Section 3 describes our three-step method for collocation identification in corpora and filtering; in Section 4 we present our evaluation procedure, and we describe an analysis of the effects of frequency, dispersion and AMs on human ratings of a subset of the selected collocations, which we discuss in Section 5.

2 The definition of collocation

In recent decades, many proposals have been made to define conventional word combinations and a wide variety of terms have been adopted to make reference to them (e.g., Evert 2009; Halliday 1961; Sinclair 1991; Wray 2002). In this study, we adopt the term *collocation* to refer to this phenomenon and we operationally define collocation by relying on specific theoretical assumptions and integrating previous definitions from Gries (2008) and Zanda (2025): a collocation is the co-occurrence of two syntagmatically related words with a grammatical configuration and a syntactic relation, commonly used in a range of texts and observed in corpora more frequently than expected on the basis of chance, where the two words are strongly associated, either adjacently or within a distance, and whose conventional meaning as a combination may be more or less compositional and more or less transparent based upon the senses of its components. Theoretically, this definition implies that a collocation is characterised by its

conventional meaning, resulting from the number of times it is used in naturally occurring language (frequency), the range of texts where it occurs (dispersion) and the extent to which its components attract each other (in terms of AMs). Depending on these measures, a collocation can be semi-transparent, with a component being used with a non-literal meaning (*run the risk*), non-compositional, if its meaning does not stem from the sum of its components (*spill the beans)*, or conversely fully transparent and compositional (*train station*, *cold weather*), if its repeated and widespread use and the strength of association of its components are high and make it conventional.

Compared to many of the previous ones, our definition attaches a relevant role to dispersion, and thus to the extent to which a collocation is spread across a variety of textual genres. This is particularly significant from the perspective of a dictionary designed for L2 learners, whose entries are also selected on the basis of how representative they are of the linguistic input to which learners are potentially exposed, which is supposed to reflect a wide variety of genres. Previous studies have already demonstrated the crucial role of dispersion on the acquisition of L2 vocabulary (e.g., Hashimoto and Egbert 2019) and specifically of multi-word units: dispersion was found to be a predictor of acquisition and productive use by learners, since more dispersed collocations are acquired earlier than domain-specific ones (Chen 2015), especially over time (Candarli 2021). Since the breadth of contexts in input affects collocational learning, studies also suggest that learner dictionaries and coursebooks should include collocations drawn from varied text types (Hoang and Crosthwaite 2024). Additionally, the present definition implies that a collocation is formed by syntactically related lexical units, and this distinguishes our approach from approaches based on n-grams (e.g., Wahl and Gries 2018). Specifically, given the rich morphology of Italian, our definition of collocation explicitly incorporates this aspect by stressing that collocations are formed by lexical items consisting of a specific syntactic relation. This choice allows us to account for the high degree of inflectional variation in Italian, ensuring that the definition does not rely solely on surface forms but rather on the underlying lexical and syntactic link between the elements. Further, the syntactic configuration used to identify collocations in corpora was also included as part of the information provided in the lexical entries of the learner dictionary. Finally, our definition implies that the two syntagmatically connected words may also be non-adjacent, preserving both their syntactic relationship and their collocational status.

3 Literature review

In this section we review previous research on the identification and extraction of candidate collocations from corpora (3.1), on the process of filtering them using

statistical measures (3.2), on the thresholds that can be used for these measures (3.3), and on the evaluation of methods used in these tasks (3.4). The main thread in this review relates to the implications that linguistic theory and language corpora have on the methodological choices involved in the extraction of collocations from corpora for the creation of a learner dictionary. These implications include the very notion of collocation as a linguistic element at the interface between lexicon and grammar, and hence the need to adopt extraction methods that take into account both lexical and syntactic features. The most relevant assumption stemming from linguistic theory, however, is that a word pair can be considered a collocation if there is a certain degree of association between its components. From a methodological point of view this assumption leads to the selection of specific AMs capable of quantifying this association. In this context, one of the key studies we relied on is Gries (2022a), who establishes some fundamental points that are also particularly relevant for this study. His main argument, which will be explored further below, is that many of the existing AMs do not really measure association between word pairs, but react more to frequency than they do to association, and thus conflate different collocational properties.

Given the aim of our work – the description and evaluation of an approach to identify collocations with lexicographic relevance – we circumscribe our review to specific studies that share a similar lexicographical focus, as well as to previous research that carried out comprehensive and comparative evaluations of large sets of statistical measures. Although these studies have not yet led to univocal and consistent results and they are not specifically targeted at Italian, it is possible to identify a core number of findings which are common to most of them, forming a consolidated body of achievements that can be applied to the context of Italian.[1]

3.1 Identification of collocations in corpora

AMs allow to estimate the magnitude and significance of the association between two words, that is, to identify how much one word tends to appear together with another

1 Most the literature presented is based on languages other than Italian. As a reviewer pointed out, studies on languages with less rich morphology than Italian may be not equally relevant to our work and suggested that we focus more extensively on techniques specifically suited to Italian. We acknowledge the point; however, we believe that morphological richness is not a determining factor in this context, as the units examined in our study are lemmas, grammatical categories, and syntactic relations. Much of the literature we review concerns English, as is often the case, and the specificity of Italian becomes relevant primarily when selecting which POS sequences to include in the dictionary, an aspect for which we cite studies that support our choices. To our knowledge, no previous work has carried out a systematic and large-scale extraction of collocations from Italian corpora for lexicographic purposes.

compared to what would be expected by chance. Therefore, AMs are generally considered as more effective in the identification of dictionary-relevant collocations than frequency alone (Dobrovoljc 2020; Evert et al. 2017). However, a large majority of the reviewed studies agree that previous research on AMs has produced inconsistent results (Deng and Liu 2022; Evert et al. 2017), and that the choice of specific AMs and extraction methods largely depends on the purpose of the task (Bartsch and Evert 2014; Su et al. 2024). Moreover, there are considerable differences in the performance of different AMs depending on combination types (Bhalla and Klimcikova 2019; Evert and Krenn 2001; Gablasova et al. 2017; Garcia et al. 2019; Su et al. 2024), corpus size (Bartsch and Evert 2014; Deng and Liu 2022; Dobrovoljc 2020), and text genres (Deng and Liu 2022; Gablasova et al. 2017). Consequently, these elements should always be considered as variables potentially influencing collocation identification and extraction. Additionally, previous research has highlighted that the morphosyntactic classes of the collocation components are relevant for lexicographical purposes (Bartsch and Evert 2014; Krek et al. 2022; Orenha-Ottaiano et al. 2021), thereby supporting the methodological choice of extracting collocations based on their syntactic configuration. This approach aligns particularly well with Italian, whose rich morphology makes morphosyntactic relations especially salient in the identification of collocations. On the effectiveness of the use of parsed data, a general consensus has emerged since the work of Seretan (2011), who demonstrated that syntactic dependencies, for their ability to capture syntactic relations between words, and thus discontinuous and flexible collocations, improve the quality of the task of candidate collocations detection (Bartsch and Evert 2014; Bhalla and Klimcikova 2019; Evert et al. 2017; Garcia et al. 2019). However, the issue of parsing error rates affecting the accuracy of the detection task has been raised (Constant et al. 2017). Once again, integrated methods have been proposed for this task, specifically with reference to Italian (Castagnoli et al. 2016), overcoming the limitations of both extraction methods, with promising results (Perri et al. 2024).

Beyond this body of common achievements, previous research produced no consensus on several other issues, such as the selection of individual measures of association, the thresholds to be used for filtering candidates, and the methods to assess the detection procedures.

Given the importance of AMs in identifying collocations in corpora, we will now discuss in detail the selection of different AMs and, then, the issue of thresholds.

3.2 Selection of AMs

As far as AMs are concerned, two of the most widely known studies are Pecina (2005, 2010), who evaluated 84 AMs in the identification of Czech collocations, finding that

MI performed the best among all the considered AMs. Evert's (2009) foundational work carried out a systematic evaluation of AMs used to extract significant word pairs from corpora, highlighting the existence of frequency bias for most of them, and demonstrating that the effectiveness of AMs is not universal but is a function of several factors such as the corpus used and the application goal.

In addition to these large-scale evaluations, other studies have considered more specific contexts. Bartsch and Evert (2014) assessed six AMs in the identification of collocations in four English corpora of different sizes and genres. The best performance is achieved by MI2; another finding which is relevant to this study is that increasing corpus size does not improve the performance of collocation extraction, especially if the larger corpora are less balanced and "clean". Evert et al. (2017) presented the results of a large-scale evaluation of 20 AMs on 13 corpora. Depending on the gold standard adopted, chi-squared, log-likelihood and MI2 yield the best results. Garcia et al. (2019) carried out a systematic evaluation of twelve AMs using three syntactic patterns (adjective-noun, verb object, and nominal compounds) in English, Portuguese, and Spanish. Interestingly, in syntax-based collocation extraction, raw frequency performs as well as AMs. More recently, Dobrovoljc (2020) reported the evaluation of six AMs in the identification of n-grams in written and spoken corpora of Slovenian, with Dice and MI, two measures with a long-standing tradition in lexicography (Chunk and Hanks 1990; Pecina 2010; Rychlý 2008), outperforming the others. Deng and Liu (2022) performed a multidimensional evaluation of the effectiveness of seven AMs in the identification of three types of collocations across five textual genres and seven corpora of different sizes. Their results show that Log Likelihood Ratio and MI3 are consistently more effective than the other five AMs across almost all genres, collocation types, and corpus sizes. Interestingly, corpus size has an influence on the accuracy of all the AMs under examination, but this effect becomes irrelevant after a corpus reaches the size of 150,000 tokens. In a recent study, Su et al. (2024) evaluated 16 AMs using the British National Corpus. Based on their results, MI3 is the AM with the best overall performance.

As can be seen, the results of these briefly summarised studies are far from consistent and do not allow for generalisations. Only very few findings seem to be common to most of the reviewed studies. For example, specific AMs are commonly recognised to show recurrent behaviours in collocation extraction: MI behaves differently from all the other commonly used AMs, being "a unique AM" (Deng and Liu 2022: 204) in favouring collocations composed of uncommon words (Gablasova et al. 2017), while directional measures like Delta P, despite having been used with interesting results in acquisitional studies (e.g. Ellis et al. 2014), usually obtain worse results than symmetric ones (Bhalla and Klimcikova 2019; Garcia et al. 2019).

As mentioned above, Gries (2022a) makes a fundamental contribution in this matter, since he provides strong evidence that part of the existing AMs do not really measure what they are intended to measure – association between words – but mainly frequency, or at least "some amalgam of a lot of frequency and a little association" (Gries 2022a: 1). This finding has major implications for this study, since the construct of association plays an important role in second language acquisition (e.g., Ellis et al. 2008; Durrant and Schmitt 2009, Siyanova-Chanturia 2015), and must be taken into great consideration also in the selection of entries for a dictionary of collocations that is intended for learners. Thus, according to Gries (2022a), two AMs that are much more affected by association than by frequency, and that therefore are much closer than others to being 'true' AMs, are MI and LogDice. This suggests that it would be appropriate "in terms of methodological validity" (Gries 2022a: 30) to make a clear separation of dimensions, and to use collocation frequency as one dimension, and one "true" AM as the other, as in Ellis et al. (2008) and Siyanova-Chanturia (2015).

3.3 Thresholds

Minimum thresholds of frequency, dispersion, and AMs for the identification and selection of candidate collocations are another aspect on which a general consensus has not been reached. In fact, the cut-off points strictly depend on the type of the specific collocations under investigation, the focus and aims of the study, as well as on the size of the reference corpus used.

According to some scholars, research on phraseology employs frequency thresholds ranging from 5 to 40 occurrences per million words in a reference corpus (Dobrovoljc 2020). This variability in threshold selection is also evident in lexicographical projects, which usually include various types of collocations. Cut-off points range from an absolute frequency of 10 occurrences, which roughly reach 0.01 per million words in a reference corpus of ca. 1 billion words, to identify 81 different syntactic types of collocations (Krek et al. 2022); to relative frequencies of 1 instance per million words to identify collocations of 9 different syntactic relations in multiple languages (Orenha-Ottaiano et al. 2021); to even 20 per million words for formulaic sequence extraction (Dobrovoljc 2020).

Moving on to other measures relevant to this study, while measuring frequency is rather intuitive, how to quantify dispersion in a corpus is still open to debate. Several dispersion measures (DMs) have been proposed (for an overview, see Gries 2020), but there are no thresholds that fit all purposes. Cut-off points for DMs are often established empirically, with each study developing an *ad hoc* approach, in accordance with its own goal(s) and dataset(s).

Differently from dispersion, common use cut-off points for the statistical identification of collocations have been proposed for a number of AMs, at least in corpus studies. For instance, it was suggested that combinations with MI ≥ 3 and t-score ≥ 2 would qualify a given word combination as a probable collocation (e.g., Hunston 2002). Similarly, a Log-likelihood (hereafter LL) value greater than or equal to 3.84 would qualify a word combination as statistically significant ($p < 0.05$), meaning that the co-occurrence of two or more words is unlikely to be due to chance.[2] However, there is limited evidence to support universal application of these thresholds (Durrant et al. 2022), especially when different AMs are unequally dependent on corpus size (Gablasova et al. 2017), emphasise various collocational properties (Brezina 2018: 66–74), and do not perform equally in terms of independency from frequency values (cf. Gries 2022a). Consequently, multiple AMs with divergent cut-off points were proposed to automatically identify collocations in corpora (cf., *inter alia*, Brezina et al. 2015; Deng and Liu 2022).

Overall, there is considerable variability in the measures and thresholds used, and it seems necessary to employ different methods on a case-by-case basis, depending on the aims of each study. For instance, in learner-oriented lexicographical applications, it may be appropriate to set cut-off points that allow the retrieval and inclusion of pedagogically relevant target collocations. In this context, thresholds may also be selected based on previous empirical findings, such as those by Durrant and Schmitt (2009), which show that non-native writers tend to underuse collocations with high MI score (> 7.00) in comparison with native writers, suggesting difficulties in the full acquisition and control of such collocational items. Hence, including these combinations by means of *ad hoc* cut-off points may help address learners' specific needs and collocational development.

3.4 Evaluation and gold standard

Previous research is also not homogeneous in the choice of the benchmark which is used to evaluate the extraction methods. Dobrovoljc (2020), for instance, employed expert raters to manually inspect the lists of top-ranked formulaic sequences extracted from corpora and to identify those with lexicographical relevance. Wahl and Gries (2018) evaluated via human ratings the performance of an algorithm for the extraction of multi-word expressions from corpora. Other studies relied on previous lexicographical resources and considered true collocations those already included in existing dictionaries (Bartsch and Evert 2014; Deng and Liu 2022; Evert

2 Critical values of LL equal to 3.84, 6.63, 10.83, and 15.13 respectively correspond to $p < 0.05$, <0.01, < 0.001, and <0.0001 levels of significance (cf. Rayson et al. 2004).

et al. 2017). Su et al. (2024) adopted an automatic filtering method of pooling and dictionary cleansing. Using this strategy, the process of filtering and validating collocations takes advantage of statistical measures and online resources without requiring human judgment.

Both these gold standard types have advantages and shortcomings. Existing dictionaries may be compiled exclusively with non-corpus-based criteria, and this may thus produce inaccurate evaluations: for instance, a set of combinations automatically identified in a corpus could be true collocations even if they are not listed in an existing dictionary. On the other hand, human judgements alone on lexicographical relevance suffer from a lack of objectivity (Atkins and Rundell 2008: 150). Again, therefore, it is potentially advantageous to use both information from previous dictionaries and expert human rating to reach an accurate evaluation of the performed task. Further, since expert raters are asked to evaluate how suitable the candidate collocations – extracted using corpus-based metrics – are for inclusion in a dictionary, corpus-based metrics might affect their evaluation. A positive effect would suggest a successful convergence between corpus-based and human metrics.

Previous research has shown that quantitative measures do, in fact, influence expert ratings. For example, Ellis et al. (2009) investigated the pedagogical validity of collocations extracted from academic corpora using MI and frequency, by asking language testers to rate, on a five-point scale, the extent to which they agreed that a phrase could be considered a collocation, that it had a cohesive meaning or function, and that it was worth teaching. The results showed a high intercorrelation among these dimensions. Moreover, since MI and frequency had been used to extract the collocations, the authors examined whether these two measures influenced the raters' evaluations. Both measures were found to be significant predictors, with MI having a stronger effect than frequency.

More recently, Naismith and Juffs (2025) explored the impact of collocation sophistication and accuracy on expert ratings that evaluated L2 written productions. Sophistication was defined in terms of the frequency of both the entire collocation and its constituent elements, while collocation accuracy was conceived as the acceptability of the full collocation. The authors found that only sophistication had a significant effect on the ratings: written texts including collocations composed of lower-frequency items received higher evaluations.

Although these previous studies did not address the lexicographical relevance of collocations, they suggest that expert ratings are influenced by corpus-based measures. This, in turn, suggests that such measures may also impact how suitable collocations are judged for inclusion in a dictionary, highlighting the potential of quantitative and statistical measures to identify collocations suitable for lexicographical purposes.

4 The present research

In the present study we describe and evaluate a corpus-based methodology for the extraction and identification of collocations for lexicographical purposes, complemented by a human and dictionary-based evaluation. A fully corpus-based methodology represents an innovation in the creation of lexicographical resources for Italian, especially considering the rarity of corpus-based lexical resources for Italian targeted at L2 learners. A corpus-based approach offers two significant strengths: on the one hand, collocations are extracted from a balanced reference corpus that is representative of various genres and linguistic registers, ensuring that collocations are drawn from diverse usage examples; on the other hand, collocations are filtered through frequency, dispersion and association measures, which allow the identification of salient collocations commonly used in a range of texts in the linguistic input. Our methodology consists of three main steps: the extraction of candidate collocations (Step A), the derivation of frequency information (Step B), and the filtering of candidates through statistical measures (Step C).

The following sections will illustrate these three steps: firstly, we describe the hybrid approach (i.e., the integration of pos-based and parsing-based extraction methods), which combines pos-tagging and dependency parsing, adopted for extracting candidates from a reference corpus (Step A); secondly, we present a method to derive frequency information based on the merging of the different extraction procedures (Step B); thirdly, we describe the filtering procedure through the combination of AMs with phrase frequency and dispersion (Step C). After these three steps, our extraction method is evaluated against both a benchmark of existing Italian collocation dictionaries (Lo Cascio 2012; Tiberii 2018) and human lexicographical ratings, seeking to answer to our first research question (RQ1): To what extent are candidate collocations extracted from the corpus attested in existing dictionaries? Finally, the effects of frequency and statistical measures on human evaluations are investigated using a regression model, aiming to address our second research question (RQ2): Do corpus-based measures affect human ratings on the acceptability of candidate collocations as entries of a learner dictionary?

4.1 Methodology

The PEK24 corpus was used as the reference corpus for the extraction of candidate collocations. It is representative of both written and spoken contemporary Italian.[3] It

3 The PEK24 corpus consists of the PEC24 corpus (Spina et al. 2025), an extension and upgrade of the Perugia corpus (Spina 2014), with the integration of the KIParla corpus, a corpus of contemporary spoken Italian made of more than 150 h of conversations (Mauri et al. 2019).

covers ten different genres: academic writing; administrative writing; literary fiction; non-fiction; school essays; newspapers; web texts; tv programs; film dialogues; and spoken texts. It is made of 144,931 texts for a total of ca. 48 million tokens, and it is pos-tagged using *TreeTagger* (Schmid 1994) trained with an *ad hoc* tagset based on a fine-grained set of 54 pos tags (Spina 2014). Given its composition, PEK24 includes a wide variety of text types and registers belonging to both written and spoken language. This makes it not only a reference corpus for Italian, but also a highly representative corpus. Indeed, it represents a large collection of Italian texts and text types of such a size that it provides accurate data on the quantitative distribution of linguistic features (Egbert et al. 2022).

Table 1 describes the PEK24 corpus and its sections.

All the word combinations in the reference corpus were considered without any pre-selection of target words or lemmas. Given that we define them as lexical combinations with a grammatical configuration and with syntactic relations between their components, in the final version collocations are presented based on their syntactic patterns. Therefore, we extracted the candidates focusing on the following six syntactic relations: i. Verb + Direct object (*vdobj*; *mantenere una promessa*, 'to keep a promise'); ii. Adjective + Noun/Noun + Adjective, the adjective is a modifier before or after a noun (*amod*; *brutta avventura*, 'bad adventure'; *tempo libero*, 'free time'); iii. Verb + Adjective, the adjective functions like an adverb by modifying the verb (*advmod1*; *stare zitto*, 'to stay quiet'); iv. Verb + Adverb, the adverb modifies the verb (*advmod2*; *fare presto*, 'to hurry up'); v. Adverb + Adjective,

Table 1: The composition of the PEK24 and its sections.

Section	Nr. of texts	Tokens
Written sections		
Academic writing	315	2,003,969
Administrative writing	194	1,914,625
Literary fiction	90	6,623,697
Non-fiction	107	3,172,781
School essays	25,137	6,989,768
Newspapers	104,433	6,902,522
Web texts	105,967	11,266,851
Spoken sections		
Tv programs	196	1,556,099
Film dialogues	116	1,107,484
Spoken texts	2,376	6,407,022
Total	**144,931**	**47,944,818**

The spoken section also included the KIParla corpus in the section Spoken texts (ca. 1 million tokens; Mauri et al. 2019).

the adjective is modified by the adverb (*advmod3*; *altamente positivo*, 'highly positive'); and vi. Noun + Noun, compounds made of two adjacent nouns (*comp*; *parco divertimenti*, 'amusement park'). The choice of these collocation types rests on several reasons: the *vdobj* relation is the most productive in Italian (Spina 2016), resulting in many collocations, often of high frequency and of particular interest since the relationship between the verb and its noun serving as direct object is flexible, and may involve the use of other lexical items within the collocation (articles, adjectives, adverbs). The other five collocation types involve the most common modifiers of the three major parts-of-speech (noun, verb and adjective) in Italian (Salvi 2013): a noun modified by an adjective (*amod*) and by another noun (*compound*), a verb modified by an adjective (*advmod1*) and by an adverb (*advmod2*), or an adjective modified by an adverb (*advmod3*).

4.1.1 Step A: extraction of candidate collocations

This step is crucial for determining the quality of the candidate collocations based on their syntactic relationship: the subsequent candidate filtering stage depends on the accuracy of the extraction process.

The two methods used for extracting candidate collocations are the pos-based method (henceforth, P-based) and the syntactic dependency-based method (henceforth, S-based). Both methods are particularly well suited to Italian, which has a rich morphological system: extracting by part of speech and by lemma makes it possible to capture all the morphological variants in which collocations occur. In addition, the use of syntactic dependencies allows us to retrieve collocations even when they are not adjacent in the surface string, an important advantage given the syntactic flexibility of Italian.

The P-based method assigns each token its pos, while the S-based approach identifies the syntactic relationships between the lexical items. Thus, the extraction of collocations depends on the pre-processing of data and the accuracy of grammatical and syntactic annotation. Both methods present advantages and disadvantages. The P-based method is positional (Evert 2005) and performs well in identifying adjacent combinations achieving a high accuracy threshold (97–98 % in Italian texts; Spina 2014). However, relying solely on pos patterns limits the ability to extract non-adjacent word pairs that are syntactically related or pairs that occur in reversed order within a sentence. Conversely, the S-based method can identify syntactic relationships between non-adjacent words because it is not constrained by the distance between the two elements. On the other hand, parsing has yet to reach high accuracy thresholds and still presents a significant error rate (Qi et al. 2020).

To leverage the strengths of both methods while overcoming their limitations, recent studies have integrated the two approaches, using them in a complementary

way, and have demonstrated greater accuracy in identifying collocations compared to using the methods individually (Castagnoli et al. 2016; Shi and Lee 2020; Simkó et al. 2017). Regarding extraction methods specifically targeted at Italian language, Perri et al. (2024) presented a hybrid methodology to extract candidate collocations of two different types of syntactic relationships in a small sample of texts, both written and spoken, drawn from a reference corpus of Italian. The results showed that the hybrid method outperforms the individual approaches in terms of recall (i.e., the proportion of true positive cases correctly identified), with a significant improvement in accuracy (i.e., the proportion of correct predictions out of all predictions made) and precision (i.e., the proportion of predicted positive cases that are actually correct) due to the implementation of negative parsing rules (i.e., rules capable of removing false positives).

The methodology tested in Perri et al. (2024) was used to extract candidate collocations from the entire PEK24 corpus, including all its sections. We started with the P-based approach to extract the six different types of word pairs (i.e., *vdobj*; *amod*; *advmod1*; *advmod2*; *advmod3*; *comp*). Corpus Workbench tool (CWB; Hardie 2012) and Corpus Query Processing system (CQP) were used to extract candidates by running seven separate queries to detect the six different syntactic patterns (see Appendix 1).[4] The S-based approach was applied to all types of selected syntactic relationships, except in the case of *comp* relation (given that compounds consist of two always-adjacent nouns, the P-based method was supposed to be more accurate). An additional criterion that had to be met was that the words forming the syntactic relationship had to comply with specific grammatical constraints, meaning they could only be nouns, adjectives, verbs, or adverbs.

The texts from the corpus were parsed using the Universal Dependencies framework for treebank annotation (UD; de Marneffe et al. 2021) in conjunction with the open-source Python library spaCy. The spaCy library processes sentences word by word and, for each word, generates a list of output objects that reflect the syntactic structure of the sentence and the syntactic relationships between its elements:

- *DepRel*: indicates the syntactic dependency relationship between the word and the main word in the sentence;
- *Form*: represents the form in which the word appears in the text;
- *Lemma*: indicates the base form of the word;
- *UposTag*: indicates the grammatical category of the word based on the pos tag scheme;
- *XPosTag*: includes additional information about the pos tag;

4 We used two separate queries to detect the amod relation as two orders are allowed in Italian: the adjective can occur before or after the modified noun. All the queries are listed in Appendix 1.

– *Head.i*: indicates the index to which the word in question is connected, that is, the syntactic head on which the word depends.

Moreover, additional syntactic rules (Perri et al. 2024) were introduced to enhance the model's accuracy and precision by reducing the number of false positives. Different functions were then implemented in the Python code. For instance, we used a function to recognise the *obj* relationship between a verb and a noun, while simultaneously verifying the pos of both elements and ensuring that the verb (*hanno*) was indeed the head of the noun (*fama*), in uses such as hanno$_{[\text{they have}]}$ fama$_{[\text{fame}]}$ mondiale$_{[\text{world-wide}]}$, found in broader contexts such as *Molto note per le proprietà minerali delle acque sono le sorgenti di nitrodi e di olmitello, le loro virtù terapeutiche* ***hanno fama*** *mondiale,* 'Well-known for the mineral properties of the waters are the nitroidi and holmitello springs, their therapeutic virtues are world-renowned'.

Our hybrid approach resulted from merging the two previous methods and including all the candidate collocations identified by both the P-based and S-based method. Our final dataset includes the candidates detected through this hybrid approach (888,427, 42.35 % of the total), as well as those identified only by the P-based approach (369,298, 17.60 % of the total), and those detected only by the S-based approach (839,870, 40.03 % of the total). In total, we extracted 2,097,595 candidate collocations. Table 2 summarises the total number of candidates drawn by method and syntactic relation.

4.1.2 Step B: counting frequency with the bag of words method

In the second step frequency information was derived from the PEK24 corpus. The different methodology used to extract candidate collocations (i.e., the P-based method, the S-based method, and the hybrid method) affected frequency counting. Specifically, if a candidate collocation was identified by only one method, its frequency perfectly matched the number of occurrences, corresponding to the total number of distinct contexts in which the combination appeared. However, if the

Table 2: The total number of extracted candidates by method and syntactic relation.

Method	vdobj	amod	compound	advmod1	advmod2	advmod3	Total
S-based	350,731	312,554	/	20,345	133,906	22,334	**839,870**
P-based	101,000	113,095	63,292	29,993	41,668	20,250	**369,298**
Hybrid	313,911	502,688	/	13,063	37,704	21,061	**888,427**
Total	**765,642**	**928,337**	**63,292**	**63,401**	**213,278**	**63,645**	**2,097,595**

candidate collocation was identified by both methods, there was a risk that the frequency would double-count the same context in which the candidate occurred.

To avoid this issue, we applied the technique of bag of words (Qader et al. 2019) which was performed only for collocations identified by the hybrid method. When a candidate collocation, such as *fare una passeggiata* ('to take a walk'), was identified by both the P-based and S-based methods, the two contexts (*Ieri ho fatto una passeggiata con Luca*, 'Yesterday I took a walk with Luca', vs. *Mi piace fare una passeggiata la domenica mattina*, 'I like taking a walk on Sunday mornings') in which the candidate collocation appeared were compared. A set was generated for each context, containing its unique words (Set A [*ieri*; *ho*; *fatto*; *una*; *passeggiata*; *con*; *Luca*] vs. Set B [*mi*; *piace*; *fare*; *una*; *passeggiata*; *la*; *domenica*; *mattina*]).

The comparison was carried out by identifying the intersection of the two sets, which involved determining the shared words. If the shared words constituted more than 50 % of the total words in either set, the two sentences were considered to belong to the same context. In this case, the frequency of the candidate collocation was not incremented. Conversely, if the overlap was 50 % or less, the sentences were deemed to represent different contexts, and the frequency of the candidate was increased by one. A manual evaluation of a small sample of 100 sentences resulted in an accuracy value of 92 %. This operation allowed us to derive the most accurate frequency possible and to proceed with the filtering step.

4.1.3 Step C: filtering candidates with measures of frequency, dispersion, and association

The third step involves computing and using frequency, dispersion, and AMs to filter the whole dataset of 2,097,595 candidate collocations. Raw and relative frequencies of the candidate collocations were computed with the method described in Step B.

As concerns DMs, we preliminarily considered Juilland's *D* (Juilland and Chang-Rodríguez 1964), Gries' *Deviation of Proportions* (*DP*), and its normalised variation DP_{norm} (Gries 2008). Despite being widely employed in lexicography (e.g., Davies and Gardner 2010; De Mauro 2000; Leech et al. 2001), Juilland's *D* has been criticised for several technical reasons, such as its dependency on the number of corpus parts, and reporting values outside its theoretical 0–1 range or which are not consistent with the actual dispersion of the item when found in only one corpus section (for more detailed accounts see Biber et al. 2016; Brezina 2018; Gries 2008, 2020). To tackle part of the deficiencies observed in Juilland's *D*, Gries (2008) proposed the dispersion measures *DP* and DP_{norm}, which are arguably conceptually easier to understand and practically easier to compute, while offering a robust measure of dispersion (Biber et al. 2016; Burch et al. 2017; Paulsen 2023). *DP* has been rapidly adopted in the field of corpus linguistics and has also gained increasing use in lexicography (e.g., Brezina

and Gablasova 2023).[5] Due to its methodological rigor and practical benefits over *Juilland's D*, we opted to use *DP*, and more specifically DP_{norm} – which is the normalised version of *DP* whose values always fall between 0 and 1 (Lijffjit and Gries 2012) – to quantify dispersion in our dataset in a straightforward way that may facilitate comparability with other studies.

As refers to AMs, we initially computed MI, MI3, LogDice, and LL, as they emphasise different characteristics of collocations. For instance, MI tends to give prominence to exclusively associated yet infrequent collocations. In an attempt to compensate for its low frequency bias, heuristic reformulations of MI have been proposed (e.g., MI3) but remain less widely used in language learning research literature (Gablasova et al. 2017).

LogDice has been introduced as a lexicographer-friendly AM (Rychlý 2008). Unlike MI, the peculiarity of LogDice is that it helps highlight "exclusive but not necessarily rare combinations" (Gablasova et al. 2017: 164). According to Rychlý (2008: 9), LogDice score has "reasonable interpretation, scales well on a different corpus size, is stable on subcorpora, and the values are in reasonable range".

The correlation matrix of all the measures is plotted in Figure 1. As can be seen, some values show high correlation (e.g., raw frequency and LL; MI and MI3), while others indicate moderate correlation (MI and LogDice) or low correlation (e.g., raw frequency and DP_{norm}, and even lower with MI).

In order to integrate different AMs and capture specific properties of collocations, we decided to use two "true" AMs (Gries 2022a), based on the following motivations:

– MI, to identify cohesive, salient and very strongly associated collocations, although potentially not frequent, using a high threshold, as we discuss in the next section;
– LogDice, to detect still exclusive and not necessarily rare collocations, removing the poorly associated ones.

Furthermore, we used LL values – and particularly the LL = 0 ones – as indicators of poor statistical significance of association. As mentioned in Section 3.3, LL focuses on the statistical significance of the observed frequency of co-occurrence of a word combination compared against the null hypothesis of independence, i.e. the expected frequencies of its elements (Dunning 1993). Consequently, combinations with LL values equal or close to 0 signal no or poor statistical evidence of association. For this

5 More recently, Gries (2022b) noted that most dispersion measures "reflect frequency more than they do dispersion" (2022b: 202), thus conflating frequency and dispersion in their output. Hence, to overcome the overlapping issue, Gries (2024) suggested the use of the information theoretic measure of Kullback–Leibler divergence or D_{KL} which, together with its variant $D_{KLnofreq}$, appears as a promising measure to be implemented in future lexicographical works.

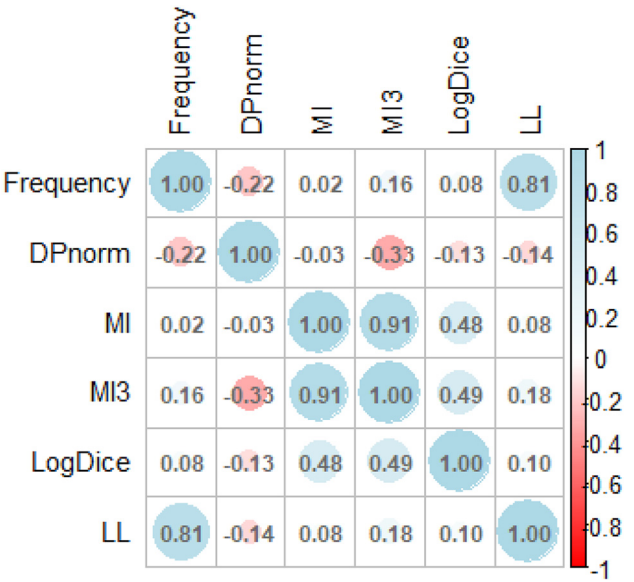

Figure 1: Correlation matrix of collocation frequency, dispersion, and association measures in the whole dataset.

reason, LL can also be employed empirically as a diagnostic tool, serving as a strong indicator that certain items may not be genuine collocations. Hence, in our approach LL was not employed as an AM with a set threshold, but rather as a heuristic device to locate statistically non-significant combinations. In most cases, these combinations proved to be problematic due to spelling or tagging errors (see Section 4.2).

The descriptive statistics of raw frequency, dispersion, and AMs of the whole dataset are shown in Table 3.

Accordingly, we developed a three-stage filtering process for candidate collocations, integrating dispersion with AMs and with frequency on the basis of their specific characteristics and statistical correlation (cf. Figure 1). Dispersion was our core measure, which we maintained across all the filtering stages, setting an empirical threshold of $DP_{norm} \leq 0.55$.[6] This approach enhances the reliability of extracted collocation lists, as sufficiently highly dispersed phenomena indicate a degree of independence from corpus composition, while low dispersion suggests context-specific occurrences. The strength of cohesion between the components of

6 DP_{norm} values close to 1 represent low dispersion, while those close to 0 indicate high dispersion. In the Frequency Dictionary of British English (Brezina and Gablasova 2023), the authors considered items with DP < 0.2 values as evenly distributed, while those with DP > 0.6 as unevenly distributed.

Table 3: Summary of the main measures of frequency, dispersion and AMs computed for the whole dataset.

Measures	Min	1st qu.	Median	Mean (SD)	3rd qu.	Max
Raw frequency	1.000	1.000	1.000	2.687 (18.55)	2.000	7,972.000
DP_{norm}	0.064	0.851	0.879	0.872 (0.14)	0.963	1.000
MI	−14.191	3.900	4.449	7.017 (4.28)	7.017	31.548
LogDice	−5.853	1.781	3.110	3.349 (2.47)	4.660	22.459
LL	0.000	0.18	4.74	13.63 (133.02)	11.63	64,435.830

the collocations is our second guiding criterion, in which we combined two different measures to sequentially capture the uncommon and highly cohesive associations and then the more frequent and still exclusive ones. Frequency was used to remove the many occasional candidates, and LL to exclude the remaining combinations without highly significant co-occurrence values, mostly due to tagging or orthographic inconsistencies. This process is described in detail in the next section.

4.2 The three-stage filtering process of Step C

Stage C1. The dataset was filtered by setting a minimum threshold for DP_{norm} (≤0.55) and MI (≥7.00) – in line with literature and previous findings (cf. Brezina and Gablasova 2023; Durrant and Schmitt 2009, respectively) – regardless of raw frequency. The overall rationale was to address learners' needs (cf. Durrant and Schmitt 2009) and include exclusively associated collocations which were at the same time scattered throughout the corpus' sections, i.e. minimum 1 occurrence in at least three different sections of the PEK24. This process allowed us to retain 16,660 collocations out of over 2 million.

Furthermore, LL was remarkably effective for identifying problematic candidate collocations featuring parsing and tagging errors – for instance, the combination *chi attivo* (in English, lit. 'who active'), incorrectly tagged as vdobj by the parser – or orthographic inconsistencies – such as non-existent forms derived by typos (e.g., *alimentere fiore* instead of *alimentare fiore* ['nourish/feed flower']: these combinations were traced back automatically (LL = 0) checked and excluded from the list. (cf. Figure 2).

Following this semi-automatic filtering process, combinations including foreign words (e.g., *green economy*) and geographical references (e.g., *origine sarda* 'Sardinian origin') were excluded through a human evaluation conducted by the authors. Finally, the number of collocations retained in Stage C1 was 12,201, resulting from the removal of 4,459 combinations.

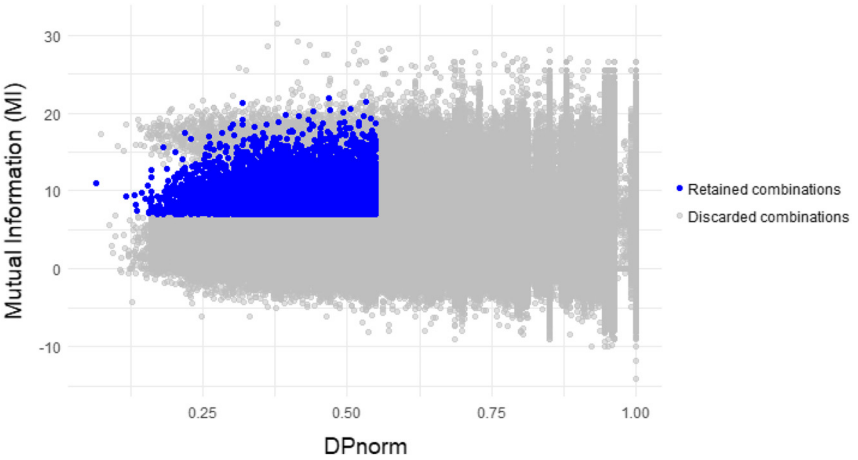

Figure 2: Stage C1: semi-automatic filtering of the full database (2,097,595 combinations). X-axis: DP_{norm} (lower = better dispersion); y-axis: MI (higher = stronger association). Each dot represents one candidate collocation. Blue dots meet thresholds of $DP_{norm} \leq 0.55$ and MI ≥ 7.0; grey are discarded. Items automatically flagged as errors (LL = 0) were manually checked and discarded.

Stage C2. Filtering involved all items from the whole database with MI < 7.00 (the remaining ones from Stage C1), with a minimum raw frequency value of 30 occurrences (i.e., 0.63 per Million words), and again $DP_{norm} \leq 0.55$. The rationale behind these thresholds is to remove occasional combinations, since in such a large dataset most of the items have very low frequencies. The total number of collocations automatically identified in Stage C2 was 9,177 (cf. Figure 3).

Stage C3. The 9,177 candidate collocations identified in Stage C2 were further filtered setting a minimum threshold of LogDice ≥ 5.00 (cf. Figure 4). In addition to the motivations asserted for the filtering in Stage C2, the idea behind the thresholds set in Stage C3 is to retain strongly enough associated collocations that are at the same time salient while not necessarily composed by infrequent words.

The resulting group of 5,346 candidate collocations was further cleaned from parsing and tagging errors using LL = 0 and manually checked to exclude foreign words and geographical references. Consequently, the final number of candidate collocations retained in Stage C3 tallied 4,619, resulting from the removal of 727.

Detailed descriptive statistics of frequency, dispersion and AMs of the resulting dataset from the three-stage filtering process are provided in Table 4.

The filtering process of 2,097,595 units produced a final list of 16,820 candidate collocations (12,201 from Stage C1 + 4,619 from Stage C3), distributed among the six syntactic types.

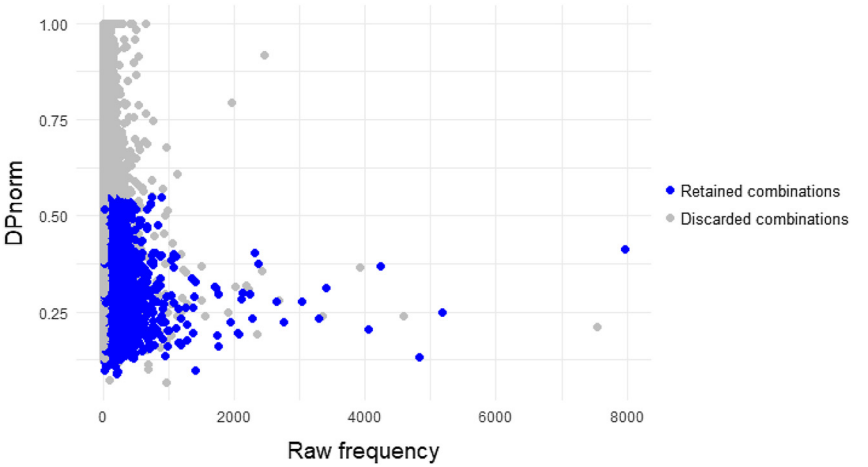

Figure 3: Stage C2 automatic filtering of all combinations with MI < 7.00 (i.e., those not selected in stage C1). X-axis: raw collocation frequency; y-axis: DP_{norm}. Blue marks meet threshold of raw frequency ≥30, $DP_{norm} \leq 0.55$); grey are discarded.

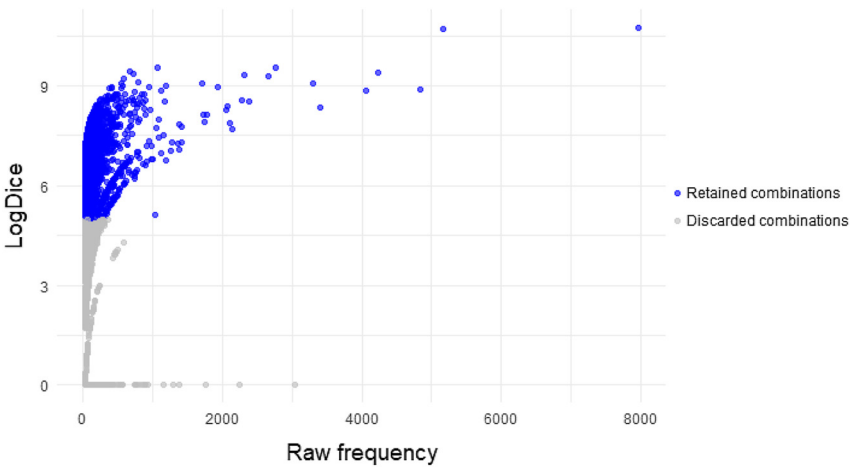

Figure 4: Stage C3: automatic filtering of the 9,177 items retained in stage C2. X-axis: raw collocation frequency; y-axis: LogDice (higher = stronger association). Blue dots indicate retained collocations (LogDice ≥ 5.00); grey are discarded.

To provide a comprehensive overview, Table 5 summarises the methodological choices of each filtering stage.

Table 4: Summary of the main measures of frequency, dispersion, and AMs for the filtered dataset.

Measures	Min	1st qu.	Median	Mean (SD)	3rd qu.	Max
Raw frequency	3.000	6.000	16.000	52.55 (172.16)	49.00	7,972.000
DP_{norm}	0.064	0.356	0.439	0.423 (0.10)	0.509	0.550
MI	0.421	6.596	7.834	7.844 (2.51)	9.200	21.456
LogDice	−0.787	5.593	6.412	6.498 (1.44)	7.313	13.742
LL	20.88	62.88	143.35	367.18 (1,275.77)	322.410	64,435.830

Table 5: Summary of the three-step candidate extraction and filtering process.

Steps	Description	Method	Output
Step A	Extracting of candidate collocations	– P-based method – S-based method – Hybrid method	2,097,595 candidates
Step B	Counting frequency of candidates extracted with the hybrid method	– Bag of words: comparisons of the contexts in which a candidate collocation occurs	Accurate frequencies for all collocation candidates
Step C	Filtering candidates with dispersion, frequency and AMs	– Three-stage filtering process; $DP_{norm} \leq 0.55$ across all stages	
		– Stage C1 retain strongly associated collocations regardless of frequency	16,660 candidates filtered
		– MI ≥ 7.00	
		– Manual and automatic check (LL = 0) to remove inconsistencies	4,459 candidates removed
		– Total retained candidates after Stage C1	12,201
		– Stage C2 remove occasional combinations	9,177 candidates filtered
		– MI < 7	
		– Raw frequency ≥30	
		– Stage C3 remove poorly associated combinations	5,346 (from the total of Step 2)
		– LogDice ≥ 5.00	
		– Manual and automatic check (LL = 0) to remove inconsistencies	727 candidates removed
		– Total retained candidates after Stage C3	4,619
Result		– **Final set of candidates retained** (Total of Stage C1 + Total of Stage C3)	**16,820**

5 Candidates' evaluation

Based on previous studies which have used existing lexicographical resources (Bartsch and Evert 2014; Deng and Liu 2022) or expert rater judgments (Dobrovoljc 2020) as benchmarks for evaluating candidate collocations, we used both types of information to establish the gold standard against which we assessed our candidate list.

As previously noted, using both information from dictionaries and expert judgments can address the challenges inherent in using either evaluation method independently. On the one hand, dictionaries, often compiled solely on the basis of non-corpus-based criteria, include a large number of combinations but may fail to capture conventional collocations. On the other hand, expert rater judgments, while potentially lacking objectivity, provide an indicator of the conventionality of combinations and thus help identify their lexicographical relevance.

In our evaluation, we addressed RQ1 (to what extent are candidate collocations extracted from the corpus attested in existing dictionaries?) by comparing the candidate list resulting from the third filtering step – based on association, frequency and dispersion measures – with two non-corpus-based Italian collocation dictionaries, which are not specifically targeted at L2 learners: Tiberii (2018) and Lo Cascio (2012). Both dictionaries were used as benchmarks because they adopt a definition of collocation that is similar to the one used in this study. They define collocations as conventional word combinations that are frequent in use and follow specific syntactic patterns. Both dictionaries, in fact, include grammatical and syntactic information that describes the types of collocations.

In Tiberii's dictionary (2018), lexical entries are organised by lemma, which belong to three grammatical categories: noun, adjective, and verb. The collocations within each lexical entry are further arranged by syntactic relation, with each relation listing the collocates of the lemma-node. For nouns (e.g., *abitudine*, 'habit'), collocations are provided for the adjective modifier relation (*buona abitudine*, 'good habit'), the Verb + Direct object relation (*abbandonare un'abitudine*, 'quitting a habit'), and the Subject + Verb relation (e.g., *l'abitudine permane*, 'the habit persists'). For adjectives (e.g., *deforme*, 'deformed'), the listed collocates include adverbs (*completamente deforme*, 'totally deformed') and verbs modified by the adjective (*rendere deforme*, 'make deformed'). Finally, for verbs (e.g., *attraversare*, 'to cross'), the collocations include adverbial modifiers (*attraversare distrattamente*, 'absent-mindedly crossing') and adjectives that modify the verb (*attraversare incolume*, 'crossing unharmed').

As in Tiberii (2018), lexical entries in Lo Cascio (2012) are organised by lemma and belong to the following parts of speech: nouns, adjectives, and verbs. Each lemma

is divided into semantic areas, with lexical combinations listed within each area. For instance, for the lemma-node *camera* (room), in the semantic area 'living space', the lexical combination *affittare una camera* ('to rent a room') is included. Within each semantic area, the combinations are further categorised by grammatical category: adjective (*camera* ***singola***, 'single room'); adverb (*affittare* ***mensilmente*** *una camera*, 'to rent a room monthly'); phrase (***camera da pranzo***, 'dining room'); noun (***camera degli ospiti***, 'guest room'); and verb (***prenotare*** *una camera*, 'to book a room').

Since the dictionaries were organised by lemmas rather than by collocations, as in our candidate list, the dictionary files were processed and reorganised to display the node on the left column and the collocate on the right. Additionally, combinations that shared the same syntactic relations as our candidates were extracted. Tiberii's list included a total of 153,942 combinations, while Lo Cascio's list comprised 52,625. Another important factor to consider was the order of the two elements in a collocation. Since two different sequences of elements (i.e., Adjective + Noun; Noun + Adjective) are allowed in Italian, to ensure the most accurate comparison possible, the match between our candidate collocations and the dictionary entries was checked in both orders.

However, neither dictionary was created based on Italian reference corpora, which may limit their coverage of the most frequent, dispersed or strongly associated collocations, as previously noted. Only Tiberii (2018) explicitly states having used an Italian corpus – Paisà – to verify the extracted collocations. However, despite its large size, Paisà is a web corpus and lacks balance in terms of text genre and linguistic register, making it less suitable for identifying collocations that are more dispersed and recurrent across a variety of genres. Precisely because the source of both dictionaries is not a representative reference corpus, we did not expect all the extracted collocations to be found in both dictionaries.

To address the issue that non-corpus-based dictionaries may include combinations that are not salient collocations, candidate combinations absent from both dictionaries were subsequently assessed by two groups of expert raters (each group made of three L1 speakers of Italian with expertise in linguistics). Raters were asked to evaluate each combination as appropriate or not appropriate for a learner dictionary according to three specific criteria:

1) Target: the dictionary is specifically designed for learners, ranging from A1 to C2 proficiency levels. As such, it should primarily include collocations of general, non-technical, and non-specialist usage (Chen 2015; Hoang and Crosthwaite 2024), unless they are technical expressions likely to have a significant impact on the life of a learner in Italy. For instance, *fotosintesi clorofilliana* ('chlorophyll photosynthesis') would likely be classified as "not appropriate", while *decreto-legge* ('decree-law') might be considered "appropriate".

2) Completeness: collocations must, on their own, without the addition of other elements, represent a unique concept. For instance, *modo chiaro* ('clear manner') would likely be categorised as "not appropriate" because it requires *in* (as in *in modo chiaro*, 'in a clear manner'). In contrast, *cronaca nera* ('crime news') is complete, as it conveys a distinct meaning on its own and can therefore be labelled as "appropriate".
3) Familiarity: collocations should resonate as familiar and conventional to L1 speakers of Italian, expressing unique concepts through a pair of words. For example, *offerta allettante* ('tempting offer') feels more familiar and conventional than *invito allettante* ('tempting invitation').

5.1 Evaluation results

The 16,820 candidate collocations extracted from the PEK24 corpus were first evaluated against the two dictionaries (Lo Cascio 2012; Tiberii 2018). The following table shows the results of the matching first against Tiberii's list, then against Lo Cascio's list, followed by the merged results (i.e., the candidate list against both dictionaries) and the total results of the comparison (i.e., candidates present in one or both dictionaries).

The comparison shows that 23 % of the candidates was found in Tiberii's dictionary and 6.8 % of the candidates was found in Lo Cascio's dictionary, while 25.2 % occurred in both dictionaries. Therefore, 59.3 % of collocations extracted from corpora occurred in the two dictionaries indicating a good alignment with the gold standard (i.e., the two dictionaries). This means that a majority of the automatically extracted collocations were attested in at least one reference dictionary, which we take as a positive indication of the method's precision. The collocations that matched were generally high-frequency and strongly associated combinations, such as *fare attenzione* ('to pay attention'), indicating that both our list and the benchmark dictionaries indeed include conventional word combinations.

However, 40.7 % of candidates were not found in either dictionary. This may be due to the fact that the two dictionaries did not adopt a corpus-based approach in identifying collocations, failing to include salient and figurative collocations *(battere le ciglia*, 'to bat one's eyelashes'), or, conversely, that the word sequences extracted from the PEK24 corpus were false positives *(libro bianco*, white book).

Therefore, the total of 6,843 combinations not found in either dictionary was divided approximately equally between two groups for evaluation (group 1: 3,419; group 2: 3,424). Of the 6,843 candidate collocations, 1,548 (22.6 %) were judged as collocations by all three raters (*bel colpo*, 'good shot'); 1,891 (27.7 %) by only two raters (*mandare un'email*, 'to send an email'); 2,221 (32.4 %) by only one rater (*gatto nero*,

'black cat'); and 1,183 (17.3 %) were marked as non-collocations by all three raters (*domenica sera*, 'Sunday evening'). Candidates that were annotated as collocations with two positive judgments out of three were reviewed by two other raters: one from group 1 checked those from group 2, and vice versa, to identify which collocations should be retained. Of these, 749 (11 %) were deemed suitable for the dictionary by the two raters. In contrast, collocations with only one positive judgment out of three, and those marked as non-collocations by all the three raters, were discarded.

Finally, the final set derived from the evaluation procedure includes 12,274 collocations: 9,977 sourced from one or both dictionaries and 2,297 from human evaluation. Overall, this integrated approach allowed us to include in our final list 72.9 % of collocations extracted from the corpus: 9,977 collocations matched with the dictionaries (59.3 %), while 13.6 % (2,297) of collocations extracted via the corpus-based method were evaluated as suitable by human raters (Table 6).

The final list of collocations drawn by extraction method and type is shown in Table 7. Collocations' examples for each syntactic relations can be found in Appendix 3. The hybrid method has outperformed both the P-based and S-based methods in extraction, with 94.8 % of collocations extracted compared to 4 % and 1.2 %, respectively. The most frequent collocation type is *amod*, accounting for 58 %, followed by *vdobj* with 31 %, while other syntactic relations are less frequent (Figure 5).

5.2 Variables affecting human rating: a statistical model

As previous studies suggest (e.g., Ellis et al. 2009; Naismith and Juffs 2025), expert ratings are affected by corpus-based metrics. Therefore, the human evaluations were further analysed using a regression model seeking to answer to RQ2: Do frequency, dispersion and association measures affect human judgements on the acceptability of candidate collocations as entries of a learner dictionary?

The dependent variable SCORE is based on the combined evaluations given by the three evaluators from the two groups, starting with the lowest (three negative evaluations = 0), and proceeding with two negative and one positive evaluations (=1), two positive and one negative evaluations (=2) and three positive evaluations (=3).

Table 6: Results of the comparison between the candidate list and the two dictionaries.

Comparisons	Total (%)
Total number of matches in Tiberii's dictionary	3,875 (23 %)
Total number of matches in Lo Cascio's dictionary	1,856 (6.8 %)
Total number of matches in both dictionaries	4,246 (25.2 %)
Total number of matches in one or both dictionaries	9,977 (59.3 %)
Total number of no-matches	6,843 (40.7 %)

Table 7: The final set of collocations drawn by extraction method and type.

Method	amod	vdobj	advmod2	advmod1	advmod3	compound	Total
S-based	124	12	16	1	/	/	**153**
P-based	83	25	148	23	40	160	**479**
Hybrid	6,883	3,761	584	217	197	/	**11,642**
Total	**7,090**	**3,798**	**748**	**241**	**237**	**160**	**12,274**

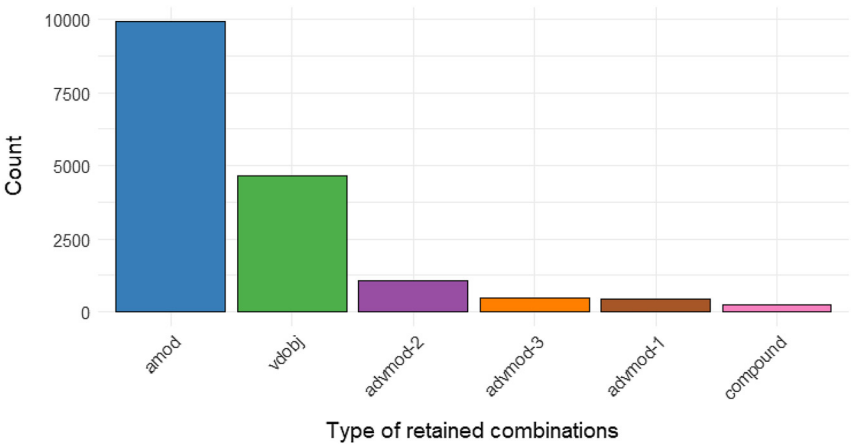

Figure 5: The final set of collocations distributed across six syntactic types.

The observations are the 6,843 candidates that were not retrieved in the dictionaries and were submitted to the assessment of human evaluators. Table 8 shows the descriptive statistics of frequency, dispersion and AMs of candidate collocations rated by human raters.

As we were not interested in a binary assessment of yes/no decisions, but rather in a scalar aggregation of the assessments of the two groups of three raters, we relied on a regression model to analyse the degree to which the different measures affected human ratings.

We used an ordinal logistic regression model, as our dependent variable SCORE was ordinal. As independent variable we entered the measures used to filter candidate collocations, namely frequency per million tokens, MI, LogDice, DP_norm and LL, and the categorical variable TYPE, assuming that the syntactic configuration of the collocations can affect human ratings as well, possibly in interaction with statistical measures. The model was built using the R software (R Core Team 2023),

Table 8: Summary of the main measures of frequency, dispersion, and AMs for candidate collocations evaluated by experts.

Measures	Min	Median	Mean (SD)	Max
Raw frequency	3.000	10.000	40.44 (121.97)	4,057.000
DP_{norm}	0.064	0.463	0.44 (0.09)	0.550
MI	0.421	7.876	7.86 (2.78)	21.456
LogDice	−0.787	6.053	6.169 (1.513)	13.742
LL	20.88	101.986	247.456 (725.308)	29,853.16

the packages MASS 7.3–55 (Venables and Ripley 2002) and effects 4.2–2 (Fox and Weisberg 2019). We adopted a top-down approach for the model selection procedure. Accordingly, we started including the maximum number of fixed effects and interactions, and we explored them to create the best model fit. In the process of comparing pairs of models, we relied on the Akaike information criterion (AIC) indicating the amount of variance that is left unexplained by the model. At the end of the model selection procedure, we checked for multicollinearity for all the fixed effects without interactions in the final models using Variance Inflation Factors (VIF): all VIF scores exhibited a low correlation (VIF = 1; James et al. 2013).

5.2.1 Model results

Results of the ordinal logistic regression model indicate that all the predictors reach standard levels of significance (see model's summary in Appendix 2). Specifically, frequency has a negative value, meaning that an increase in frequency increases the likelihood of a negative human rating (scores = 0, 1). Conversely, MI has a positive value: an increase in MI, and thus in how strongly and exclusively the words included in collocations are associated, increases the likelihood of positive human evaluations (scores 2, 3). Dispersion (DP_norm) and LogDice have significant interactions with the syntactic configuration of the collocations (TYPE): they both affect human ratings to different extents as a function of collocation types.

Figures 6–10 visually represent these results: in Figure 6, the plot shows that positive evaluations (score = 3) are directly correlated with a linear increase in MI, and that, conversely, negative ratings (score = 0,1) are more likely with lower MI values. Something similar happens for LL (Figure 7), where high values affect particularly score = 3. Conversely, a high frequency per million tokens (Figure 8) predicts the likelihood of the most negative human evaluation (score = 0), while positive evaluations (score = 2,3) tend to be more likely predicted by lower frequency values. As already mentioned, the two other AMs show significant interactions with

collocation type. The effect of dispersion (Figure 9) appears to be symmetrical between positive evaluations (score = 3), which are predicted by a higher dispersion (the values of DP_norm are in inverse order, meaning that a higher dispersion, corresponding to a lower value, predicts higher human ratings), and negative evaluations (scores = 0,1), where dispersion values are lower. This positive effect of dispersion on human scores has different degrees depending on collocation types: verbs modified by adjectives (advmod-1), for example, have the highest probability to predict positive human evaluations (>50 %) for highly dispersed collocations, followed by verb + direct object (vdobj) collocations (40 %). Similarly, high LogDice values (Figure 10) predict positive human ratings (score = 3), particularly for advmod-1 collocations, that reach a likelihood of a positive evaluation >60 %. The interaction between LogDice and the advmod-3 collocation type is the only one that does not reach significance. For negative human ratings, the advmod-1 and the amod collocation types are the ones where the impact of low LogDice values is stronger on the probability to get scores = 0,1.

The model has a weak effect: the Nagelkerke's R2 value, which quantifies the amount of variance explained by the predictors, is 0.083, although the model includes 6 independent variables. However, it is acknowledged that in logistic regression R2 is difficult to measure, and that an equivalent statistic to R-squared actually does not

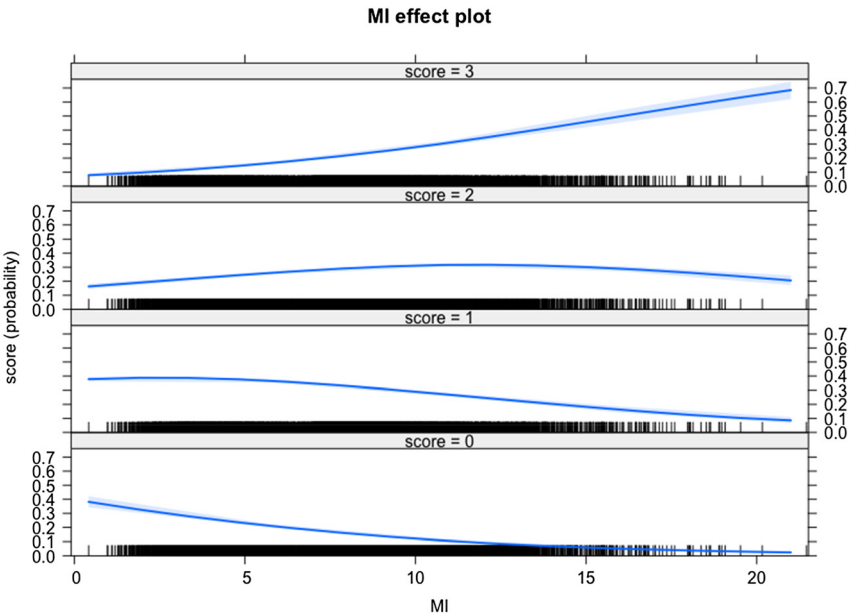

Figure 6: The effect of MI on human ratings.

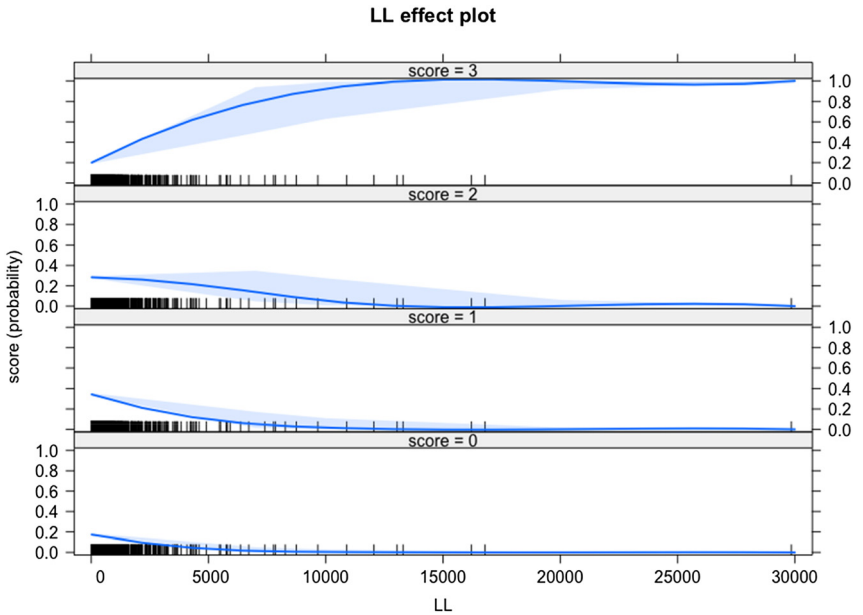

Figure 7: The effect of LL on human ratings.

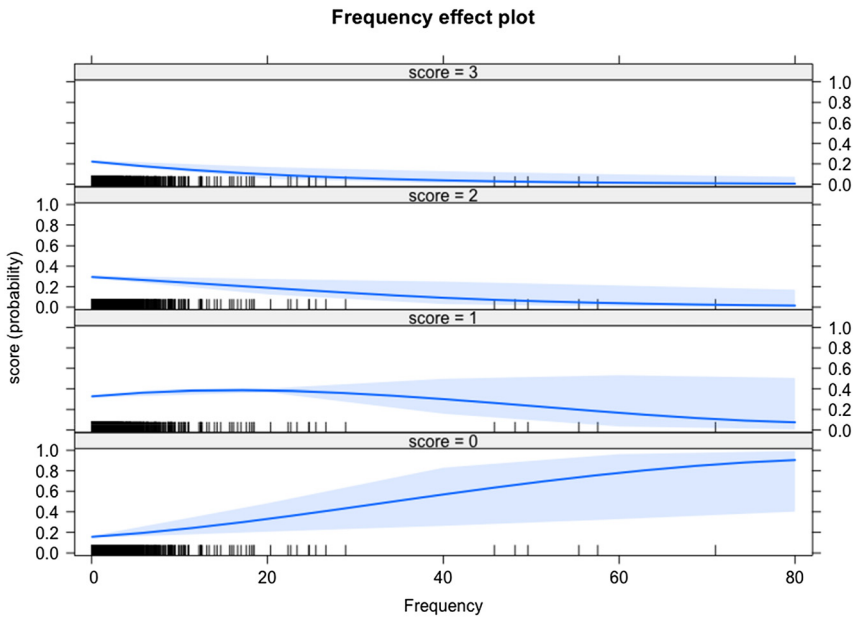

Figure 8: The effect of frequency on human ratings.

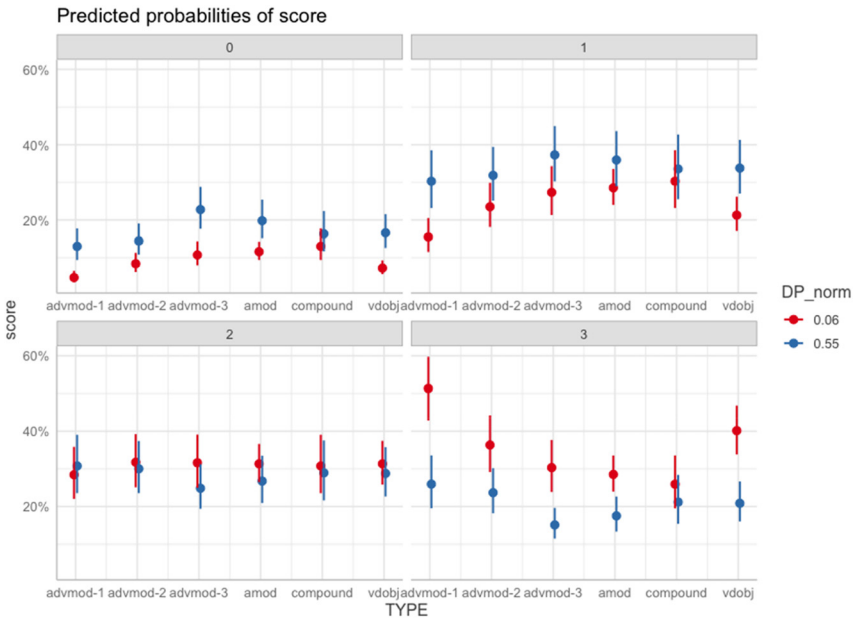

Figure 9: The interaction of the effects of dispersion and collocation type on human ratings.

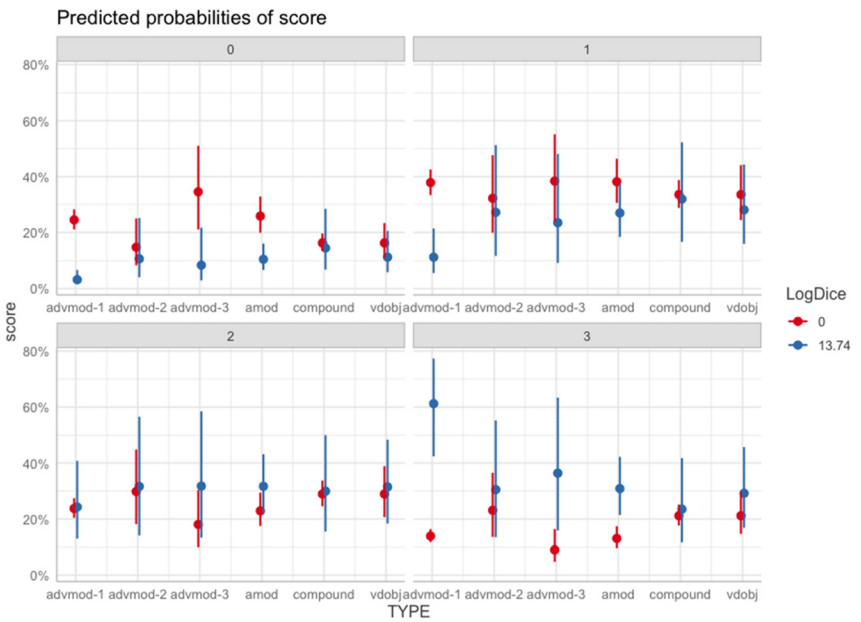

Figure 10: The interaction of the effects of LogDice and collocation type on human ratings.

exist (Baguley 2012). The fairly large sample size (n = 6,843) and the all-significant predictors in any case indicate rather clear trends in the correlation between the measures considered and human evaluations.

6 Discussion

In the previous sections we described a procedure for the identification, extraction and filtering of collocations from corpora. This procedure is articulated in different steps and characterised by the integration of different approaches (S-based, P-based, hybrid) and of different measures (dispersion, frequency, LL, MI, LogDice). We then evaluated this procedure, once again by implementing different methods: we first matched our candidate list against two existing dictionaries, and then we submitted the non-matching candidates to the assessment of six human raters. This procedure allowed us to identify 12,274 final collocations from an initial candidate set of 2,097,595. We finally investigated if and to what extent the statistical measures involved in the different steps affect human raters, as well as their potential interactions with the collocation types.

In the first step (Step A) of our procedure, we extracted candidate collocations that matched six syntactic relations by integrating the two traditionally used methods of pos-tagging and dependency parsing. Our previous study on a sample of a reference corpus (Perri et al. 2024) already demonstrated the effectiveness of a hybrid method, which was able to outperform the individual P-based and S-based approaches in terms of recall and benchmark match. The integration of hybrid and individual approaches, extended to the entire corpus, yielded a double gain: on the one hand, we were able to identify and include in the final dataset of selected collocations 632 (5.2 %) collocations individually extracted through pos-tagging (n = 479) and dependency parsing (n = 153), part of which would not have been identified if we only had relied on an individual approach; on the other hand, the vast majority (94.8 %) of the collocations included in the final dataset were identified through a hybrid procedure, both P-based and S-based. This implies that the coefficients underlying the subsequent filtering procedures (frequency, dispersion and AMs) were calculated using more accurate occurrences values of the collocations, which also include non-adjacent or inverted occurrences in the order of their components. The third step of our procedure (Step C) was devoted to the filtering of the 2,097,595 candidate collocations extracted from the reference corpus. To this aim, we have sequentially applied several quantitative measures that previous research had recognised as effective in different ways and to different extents depending on the research contexts. Preliminarily, we chose to use dispersion as the core measure able to include word combinations that are widespread across a range of text genres

and to exclude those that are typical of only specific text types. Previous research has shown the crucial role of dispersion on the L2 acquisition of multi-word units, where more dispersed collocations are acquired earlier than domain-specific ones. Therefore, we set a cut-off value for dispersion (DP_norm ≤ 0.55), and we kept it constant throughout the successive stages of the filtering procedure. The analysis of the effect of the different measures on the human ratings of 6,843 candidates has proved that this choice was effective: an ordinal logistic regression model indicated that among all the coefficients used, dispersion was the one with the highest effect on human evaluation: the higher the dispersion value, the more likely is a positive rating from a human evaluator. This finding is even more significant if one considers that frequency has an inverse effect: a higher frequency of a candidate increases the likelihood of a negative evaluation by the rater. The model thus seems to suggest that the conventionality of a collocation, as it is perceived by human raters, derives more from how much the combination is used in different contexts rather than from its mere frequency.

In the first stage of this filtering procedure (Stage C1) we used MI, a measure traditionally used in lexicography as well as in other fields to rank collocations. We employed a high cut-off value (MI = 7), without a complementary threshold involving frequency, with the aim of capturing all the strongly associated combinations, regardless of their frequency. The regression model on human evaluation suggests that this choice was motivated: in fact, it indicates a positive effect of MI and a negative effect of frequency on human ratings: candidates with high MI and low frequency have a high probability of being assessed as acceptable for inclusion in the dictionary. This high degree of acceptability of candidates with high MI and low frequency also involves the combinations assessed as collocations through comparison with dictionaries: overall, in the final dataset the collocations with MI ≥ 7 and frequency <30 (thus at this first stage based only on MI, since the threshold of 30 for frequency was set in the following step) are 6,856 (5,351 from dictionaries matching and 1,505 from human evaluation, and 55.8 % of the total number of accepted collocations).

Similar considerations can be applied to the last stage of the procedure (Stage C3), in which LogDice was used to detect not necessarily rare but strongly associated collocations, filtering out candidates with a value of ≤5. LogDice was found to be another significant predictor in the regression model: with higher LogDice values, a positive score was more likely to be assigned by human evaluators. Therefore, establishing a cut-off point that excludes candidates with a low LogDice value may result in a set of candidates that are more likely to get positive scores from the evaluators and thus be accepted as dictionary entries. Overall, the final number of selected collocations with LogDice ≥ 5 and MI < 7, thus filtered only according to the

LogDice value above the cut-off point, is 3,333 (2,835 from dictionaries matching and 498 from human evaluation, and 27,2 % of the total number of accepted collocations).

What emerges from these considerations suggests that, in the case of collocations to be extracted from corpora on the basis of syntactic relations, detection and filtering procedures carried out in successive steps and characterised by the integration of different methods and measures may lead to a good degree of accuracy. The successive integration of dispersion, MI, frequency, and LogDice allows candidates to be filtered on the basis of the different dimensions that characterise them and affect their perception as conventional combinations: their spread across a range of genres, their repeated use, and their exclusivity, both in the case of rare (MI) and more frequent combinations (LogDice).

The results of the regression model provided interesting insights into the effects of quantitative measures on human evaluation of the conventionality of collocations. We have already mentioned the homogeneous behaviour of dispersion and of all the AMs used, which have the effect of linearly increasing the probability of positive evaluations as their value increases, whereas frequency has the opposite effect. Moreover, the results highlight different behaviours for the six syntactic configurations of collocations in their interaction with two specific measures: dispersion and LogDice. In particular, the advmod-1 relation, in which an adjective functions as an adverb by modifying the verb (*stare zitto*, 'to stay quiet'), is the one with the highest probability of resulting in a positive human evaluation when associated with a high dispersion value or with a high exclusivity value (LogDice). This result is all the more relevant because advmod-1 collocations are few in number compared to other types: in our final dataset of 12,274 collocations, they are only 241 (2 % of the total). In fact, this is a small set of very common, dispersed and, most importantly, highly familiar collocations, which are therefore easily perceived as conventional by human evaluators. A few examples of advmod-1 collocations may better clarify this point: *stare buoni* ('to be good'), *battere forte* ('to beat hard)', *mangiare sano* ('to eat healthy'), *spararla grossa* ('to talk big'). Our evaluation procedure allowed us to effectively identify many of these collocations, even those that, probably due to their colloquiality, had not turned up in the comparison with the existing dictionaries. The regression model, moreover, has captured their significant interaction with the measures of dispersion and association and their combined effect on human evaluation. These results highlight the potential of corpus-based measures to identify collocations suitable for lexicographical purposes.

As a final point, we would like to reflect on the impact of selecting a reference corpus. In the case of PEK24, we are dealing not only with a reference corpus of Italian, but also with a corpus characterised by high representativeness. As Egbert et al. (2022: 11) define it, representativeness is "the extent to which a corpus permits accurate generalizations about the quantitative linguistic patterns that are typical in

a target language or discourse domain." Representativeness involves two considerations: the extent to which the collected texts reflect the range of text types and registers in the language of interest (*domain considerations*), and the extent to which the corpus accurately represents the quantitative distribution of the linguistic features under investigation (*distribution considerations*) (Egbert et al. 2022).

On the one hand, PEK24 includes different textual genres and linguistic registers not only of written but also of spoken Italian, all qualitatively documented and defined through precise metadata. On the other hand, its sections might approximate a representative distribution of the linguistic features of interest (i.e., collocations). However, as Egbert et al. (2022) rightly note, it is important to verify whether the size of each subcorpus is statistically sufficient to allow generalisations about the distributions observed within it. This remains a challenge that can be addressed in future iterations of PEK24. Nonetheless, the corpus already provides a detailed qualitative account of the texts included and of their main characteristics (Spina et al. 2025).

To address the lack of a systematic evaluation of subcorpus size and to extract collocations in a balanced way, we applied a DPnorm threshold of ≥0.55 at all stages of the extraction process. This approach enabled us to extract collocations appearing across a range of genres and registers. For instance, the collocation *mettere il punto* ('to put an end to') is highly dispersed (DPnorm = 0.88), occurring in 9 out of 10 corpus sections; by contrast, *svuotare il frigo* ('to empty the fridge') has lower dispersion (DPnorm = 0.55), occurring in only 3 out of 10 sections, which illustrates its association with specific genres and registers.

Using a highly representative reference corpus therefore makes it possible to extract not only widely dispersed collocations but also those that are more genre- and register-specific, thereby approximating the actual distribution of this lexical phenomenon in the language under study.

7 Conclusions

In this study, we described and evaluated a procedure for the identification and extraction of collocations from corpora, with the aim of compiling a Learner Dictionary of Italian Collocations. This procedure allowed us to identify 12,274 final collocations from an initial candidate set of 2,097,595. The final list resulted from an evaluation task that assessed 72.9 % of the combinations automatically filtered through frequency, dispersion and AMs (16,820) as suitable for inclusion in a learner dictionary. Importantly, through the evaluation of corpus-based methods with a

human rating task we were able to include in the dictionary 2,297 more collocations than those already attested in dictionaries (13.6 %).

Theoretically, we relied on a frequentist approach and considered collocations word combinations with a conventional meaning, resulting from their frequency in naturally occurring language, from their dispersion in texts of different genres, and from the extent to which their components attract each other. We also relied on a grammar-based approach, in which collocations are co-occurrences of words belonging to specific parts-of-speech connected by syntactic relations.

Then we translated our theoretical notion of collocations into concrete measures and brought out their conventional dimension through different values. Dispersion was the core measure in our approach, and was applied at the earliest stage with the same cut-off point to all candidates. This choice builds on the results of previous research, which showed that more dispersed collocations are acquired earlier by L2 learners than domain-specific ones. MI was used with a high threshold and allowed us to identify a large number of rare and exclusive collocations. Frequency was used with a relatively low threshold, which, in combination with LogDice, still made it possible to discard a very large number of occasional and non-conventional candidates.

We then evaluated to what extent our list of candidate collocations extracted from a reference corpus are attested in existing dictionaries. Our finding was that the majority of the automatically extracted collocations were attested in at least one reference dictionary, however leaving a substantial quantity of collocations absent from both dictionaries. These collocations were subsequently assessed by two groups of expert raters. Building on findings from previous studies indicating an influence of corpus-based metrics on human ratings, we further analysed the collected human evaluations using a regression model. Frequency, dispersion and AMs were found to influence human judgements on the acceptability of candidate collocations as entries of a learner dictionary. The measure of dispersion, together with AMs, revealed the highest positive effect on human ratings of conventionality, in interaction with collocation type. This finding suggests that the conventionality of a collocation, as it is perceived by human evaluators, derives mostly from how widespread a collocation is across a wide range of texts, in addition to how strongly its components are associated.

This study also has some limitations. First, we drew on theory and results of previous research on corpus-based approaches to collocations for the compilation of a learner dictionary, and we translated these indications into a method consisting of the integration of different measures and procedures across several steps. We are, however, aware that this does not prove that this method works better than others in

any context, and that further research is needed to systematically compare different methods, measures, collocation types, corpora and languages. Secondly, even considering only our lexicographical context, resorting to corpus-based measures of frequency, dispersion and strength of association is not definitive evidence that collocations may exist that are not attested in a corpus but are nevertheless suitable for inclusion in a learner dictionary. However, we believe that this study offers an important contribution to the field of collocation studies by showing that the triangulation of corpus-based and statistical methods, human ratings and comparison with existing resources is a crucial strategy for the effectiveness of the identification of collocations in corpora.

Appendix 1

The seven queries used to detect candidate collocations in pos-tagged data via CQP (the relation AMOD has two different queries since the adjective can precede or follow the noun).

1) VDOBJ

```
[pos="VER:.*" & lemma!="essere|andare|stare|venire|arrivare|sembrare|diventare|divenire|riuscire|rimanere|entrare|uscire|succedere|restare|piacere|morire|nascere"][pos="ADV.*" & word!="come"]?[pos="ART"]?[pos="ART|ADJ|DET.*|NUM"]?[pos="NOUN"]
```

This query excludes a list of the most common intransitive verbs and includes different potential internal configurations of the relation verb + direct object.

2) AMOD-1 (noun + adjective)

```
[pos="NOUN"][pos="ADV.*" & word!="come"]?[pos="ADJ"]
```

This query allows for the presence of an adverb between the noun and the adjective, with the exception of *come* "like" that is not a modifier.

3) AMOD-2 (adjective + noun)

```
[pos="ADJ"][pos="NOUN"]
```

4) ADVMOD-1 (verb + adjective)

```
[pos="VER.*" & lemma!="essere"][pos="ADV.*" & lemma!="come"]?[pos="ADJ"]
```

5) ADVMOD-2 (verb + adverb)

```
[pos="VER:.*" & lemma!="essere"][pos="ADV.*" & lemma!="molto|più|anche|spesso|come|oggi|sempre|domani|così|troppo|poi|meno|subito|ora|stasera|appena|già|ancora|proprio|mai |davvero|quanto|finora"]
```

This query excludes the verb *essere* "to be" and a list of the most common adverbs that can combine with any verb and do not form conventional collocations (e.g., *molto* "much", *più* "more" or common temporal adverbs like *sempre* "always").

```
6) ADVMOD-3 (adverb + adjective)
[pos="ADV.*"  &  lemma!="molto|più|meno|parecchio|sempre|anche|abbastanza|così|troppo|quasi|ancora|come|già|un_po'|tanto']
[pos="ADJ"]
```

This query excludes a list of the most common adverbs that can combine with any adjective and do not form conventional collocations (e.g., *molto* "very").

```
7) COMPOUND
[pos="NOUN"][word="-"]?[pos="NOUN"]
```

Appendix 2

The final ordinal logistic regression model with the R code.

```
m1 <- polr(score  Frequency + MI + TYPE*LogDice + TYPE*DP_norm + LL + TYPE, data = a, Hess = TRUE).
```

Model results

	Value	Std. Error	t value	CI	p-value
Frequency	−0.05	0.02	−2.89	−0.09, −0.01	**0.004**
MI	0.16	0.01	14.47	0.14, 0.18	**<0.001**
LL	3.94e−04	1.05e−04	3.74	0.00, 0.00	**<0.001**
TYPE [advmod-2] × LogDice	−0.14	0.05	−2.68		**0.007**
TYPE [advmod-2] × LogDice	−0.14	0.05	−2.68		**0.007**
TYPE [advmod-3] × LogDice	−0.04	0.05	−0.75		0.453
TYPE [amod] × LogDice	−0.09	0.03	−2.83		**0.005**
TYPE [compound] × LogDice	−0.16	0.03	−5.67		**<0.001**
TYPE [vdobj] × LogDice	−0.13	0.04	−3.33		**<0.001**
TYPE [advmod-2] × DP norm	1.00	0.15	6.70		**<0.001**
TYPE [advmod-3] × DP norm	0.42	0.16	2.73		**0.006**
TYPE [amod] × DP norm	0.96	0.14	6.71		**<0.001**
TYPE [compound] × DP norm	1.71	0.02	82.87		**<0.001**
TYPE [vdobj] × DP norm	0.34	0.11	3.15		**0.002**

Appendix 3

A sample of the resulting list of collocations for each syntactic relation.

Amod (Noun + Adjective/Adjective + Noun)

centro commerciale ('shopping mall'); *tempo libero* ('free time'); *ricerca scientifica* ('scientific research'); *capelli lunghi* ('long hair'); *attività fisica* ('physical activity'); *momento importante* ('important moment'); *strumento musicale* ('musical instrument'); *aria fresca* ('fresh air'); *lingua straniera* ('foreign language'); *punto debole* ('weak point'); *breve periodo* ('short period'); *unica soluzione* ('single solution'); *lungo viaggio* ('long trip'); *grande passo* ('big step'); *dura prova* ('tough challenge'); *stretto contatto* ('close contact'); *brutta figura* ('bad impression'); *grande città* ('big city'); *doppia fila* ('double parking'); *terribile incidente* ('terrible accident').

Vdobj (Verb + Direct object)

svolgere un'attività ('to carry on activities'); *avere l'influenza* ('to have the flu'); *dare forza* ('to give strength'); *vedere un film* ('to watch a movie'); *trovare una soluzione* ('to find a solution'); *perdere il lavoro* ('to lose one's job'); *vincere un premio* ('to win a prize'); *fornire un'informazione* ('to provide information'); *passare il tempo* ('to spend time'); *ascoltare musica* ('to listen to music'); *cambiare idea* ('to change one's mind'); *parlare una lingua* ('to speak a language'); *leggere il giornale* ('to read the newspaper'); *spendere soldi* ('to spend money'); *chiedere scusa* ('to say sorry'); *prendere una decisione* ('to make a decision'); *affrontare un problema* ('to face a problem'); *alzare la mano* ('to raise one's hand'); *porre una domanda* ('to ask a question'); *mantenere una promessa* ('to keep a promise').

Compound

fine settimana ('weekend'); *conferenza stampa* ('press conference'); *sito web* ('website'); *linea guida* ('guideline'); *parola chiave* ('keyword'); *anno luce* ('light-year'); *banca dato* ('database'); *forza lavoro* ('workforce'); *serie tv* ('TV series'); *sala giochi* ('arcade'); *paese membro* ('member country'); *pagina web* ('web page'); *punto vendita* ('point of sale'); *ufficio stampa* ('press office'); *piano terra* ('ground floor'); *lingua madre* ('mother tongue'); *fine stagione* ('end of season'); *effetto domino* ('domino effect'); *rimborso spesa* ('expense reimbursement'); *pausa pranzo* ('lunch break').

Advmod1 (Verb + Adjective)

stare zitto ('to stay quiet'); *rivelare infondato* ('reveal unfounded'); *rendere irriconoscibile* ('make unrecognizable'); *sembrare evidente* ('seem obvious'); *considerare pericoloso* ('consider dangerous'); *uscire sconfitto* ('come out defeated'); *rimanere fermo* ('remain still'); *attendere fiducioso* ('wait hopefully'); *ritenere utile* ('deem useful'); *superare indenne* ('get through unscathed'); *rendere noto* ('make known'); *risultare disperso* ('turn out missing'); *diventare amico* ('become friends'); *sembrare strano* ('seem strange'); *rimanere uguale* ('remain the same'); *ritenere giusto* ('consider fair'); *lasciare perplesso* ('leave puzzled'); *scorrere lento* ('flow slowly'); *rendere facile* ('make easy'); *diventare virale* ('go viral').

Advmod2 (Verb + Adverb)

leggere attentamente ('read carefully'); *mettere insieme* ('put together'); *studiare attentamente* ('study carefully'); *mandare avanti* ('move forward'); *cambiare radicalmente* ('change radically'); *tornare tardi* ('come back late'); *credere ciecamente* ('believe blindly'); *buttare giù* ('knock down'); *sopportare a fatica* ('bare with difficulty'); *arrivare tardi* ('arrive late'); *cacciare fuori* ('kick out'); *andare fuori* ('go out'); *ringraziare calorosamente* ('thank warmly'); *rimandare indietro* ('send back'); *uscire presto* ('leave early'); *piangersi addosso* ('feel sorry for oneself'); *parlare a vanvera* ('talk nonsense'); *passare oltre* ('move on'); *vivere bene* ('live well'); *scappare via* ('run away').

Advmod3 (Adverb + Adjective)

alquanto strano ('rather strange'); *umanamente possibile* ('humanly possible'); *potenzialmente infinito* ('potentially infinite'); *poco raccomandabile* ('not advisable'); *nettamente contrario* ('clearly opposed'); *particolarmente difficile* ('particularly difficult'); *tecnicamente valido* ('technically valid'); *indissolubilmente legato* ('inextricably linked'); *particolarmente importante* ('particularly important'); *apparentemente banale* ('apparently trivial'); *altrettanto importante* ('equally important'); *pienamente convinto* ('fully convinced'); *profondamente grato* ('deeply grateful'); *comunemente noto* ('commonly known'); *veramente felice* ('truly happy'); *pienamente consapevole* ('fully aware'); *pienamente responsabile* ('fully responsible'); *profondamente diverso* ('profoundly different'); *pressoché impossibile* ('almost impossible'); *estremamente difficile* ('extremely difficult').

Acknowledgments: The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions. The article is the result of a collaboration among all authors. Specifically, SS, IF and FZ revised all sections of all drafts of the paper. SS wrote Sections 1, 2, 3, 3.1, 3.2, 5.2, 5.2.1, 6, and 7; IF wrote Sections 4, 4.1, 4.1.1, 5, and 5.1; FZ wrote Sections 3.3, 3.4, 4.1.3, and 4.2; DP and OG wrote Sections 4.1.2. LF contributed to revising the first draft of the paper.
Research funding: This work was supported by Italian Ministry of University and Research under grant PRIN 2022HZR5E.

References

Atkins, B. T. Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Baguley, Thomas. 2012. Pseudo-R2 and related measures. In *Online supplement 4 to serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave Macmillan.

Ballance, Oliver James. 2022. Methodological considerations for the use of mutual information: Examining the role of context in collocation research. *Research Methods in Applied Linguistics* 1(3). 100024.

Bartsch, Sabine & Stefan Evert. 2014. Towards a Firthian notion of collocation. In Andrea Abel & Lothar Lemnitzer (eds.), *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern, OPAL – Online publizierte Arbeiten zur Linguistik 2/2014*, 48–61. Mannheim: Institut für Deutsche Sprache.

Bhalla, Vishal & Klara Klimcikova. 2019. Evaluation of automatic collocation extraction methods for language learning. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán & Torsten Zesch (eds.), *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, 264–274. Florence, Italy: Association for Computational Linguistics.

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4). 439–464.

Brezina, Václav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

Brezina, Václav & Dana Gablasova. 2023. *A frequency dictionary of British English: Core vocabulary and exercises for learners*, 1st edn. London: Routledge.

Brezina, Václav, Tony McEnery & Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2). 139–173.

Burch, Brent, Jesse Egbert & Biber Douglas. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 189–216.

Candarli, Duygu. 2021. A longitudinal study of multiword constructions in L2 academic writing: The effects of frequency and dispersions. *Reading and Writing* 34. 1191–1223.

Castagnoli, Sara, Gianluca E. Lebani, Alessandro Lenci, Francesca Masini, Malvina Nissim & Lucia C. Passaro. 2016. Pos-patterns or syntax? Comparing methods for extracting word combinations. In Gloria Corpas Pastor (ed.), *Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives*, 116–128. Geneve: Tradulex.

Chen, Alvin Cheng-Hsien. 2015. Acquisition of L2 collocation competence: A corpus analysis of exclusivity, directionality, dispersion and novel usage. *Taiwan Journal of TESOL* 18(1). 29–61.

Chunk, Kenneth & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Survey: Multiword expression processing: A survey. *Computational Linguistics* 43(4). 837–892.

Davies, Mark & Dee Gardner. 2010. *A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists*. London: Routledge.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308.

De Mauro, Tullio. 2000. *GRADIT: Grande dizionario italiano dell'uso*. Torino: UTET.

Deng, Yaochen & Dilin Liu. 2022. A multi-dimensional comparison of the effectiveness and efficiency of association measures in collocation extraction. *International Journal of Corpus Linguistics* 27(2). 191–219.

Dobrovoljc, Kaja. 2020. Identifying dictionary-relevant formulaic sequences in written and spoken corpora. *International Journal of Lexicography* 33(4). 417–442.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61–74. https://aclanthology.org/J93-1003.

Durrant, Philip & Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47(2). 157–177.

Durrant, Philip, Anna Siyanova-Chanturia, Kremmel Benjamin & Sonbul Suhad. 2022. *Research methods in vocabulary studies*. Amsterdam: John Benjamins.

Egbert, Jesse, Douglas Biber & Bethany Gray. 2022. *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge: Cambridge University Press.

Ellis, Nick C., Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 41(3). 375–396.

Ellis, Nick, Rita Simpson-Vlach & Carson Maynard. 2009. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *Tesol Quarterly* 42(3). 375–396.

Ellis, Nick C., Matthew B. O'Donnell & Ute Römer. 2014. Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism* 4(4). 405–431.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. Stuttgart, Germany: University of Stuttgart PhD dissertation.

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *An international handbook*, vol. 2, 1212–1248. Berlin & New York: De Gruyter Mouton.

Evert, Stefan & Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th annual meeting of the association for computational linguistics*, 188–195. Toulouse, France: Association for Computational Linguistics.

Evert, Stefan, Peter Uhrig, Sabine Bartsch & Tobias Proisl. 2017. E-VIEW-Alation – A large-scale evaluation study of association measures for collocation identification. In Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek & Vít Baisa (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference*, 531–549. Leiden, NL: Lexical Computing.

Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*, 3rd edn. Thousand Oaks, CA: Sage.

Gablasova, Dana & Vaclav Brezina. 2025. Adjective + noun collocations in L2 spoken English: How robust is the role of proficiency? *International Journal of Learner Corpus Research* 11(1). 79–113.

Gablasova, Dana, Vaclav Brezina & Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67(S1). 155–179.

Garcia, Marcos, Marcos García Salido & Margarita Alonso-Ramos. 2019. A comparison of statistical association measures for identifying dependency-based collocations in various languages. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019)*, 49–59. Florence, Italy: Association for Computational Linguistics.

Gries, Stefan Th. 2008. Phraseology and linguistic theory: A brief survey. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*, 3–25. Amsterdam: John Benjamins.

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*. Cham: Springer.

Gries, Stefan Th. 2022a. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5(1). 1–33.

Gries, Stefan Th. 2022b. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205.

Gries, Stefan Th. 2024. Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures. In *Studies in corpus linguistics*, vol. 115. Amsterdam: John Benjamins.

Halliday, M. A. K. 1961. Categories of the theory of grammar. *Word* 17(2). 241–292.

Hardie, Andrew. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3). 380–409.

Hashimoto, Brett J. & Jesse Egbert. 2019. More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning* 69(4). 839–872.

Hoang, Hien & Peter Crosthwaite. 2024. A comparative analysis of multiword units in the reading and listening input of English textbooks. *System* 121(2). 103224.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

James, Gareth, Daniela Witten, Trevor Hastie & Robert Tibshirani (eds.). 2013. *An introduction to statistical learning: With applications in R*. New York: Springer.

Juilland, Alphonse & E. Chang-Rodriguez. 1964. *Frequency dictionary of Spanish words*. The Hague: Mouton & Co.

Krek, Simon, Polona Gantar & Iztok Kosem. 2022. Extraction of collocations from the gigafida 2.1 corpus of Slovene. In Annette Klosa-Kückelhaus, Stefan Engelberg, Christine Möhrs & Petra Storjohann (eds.), *Dictionaries and society: Proceedings of the XX EURALEX international congress*, 240–252. Mannheim: IDS-Verlag.

Leech, Geoffrey, Paul Rayson & Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. London and New York: Routledge.

Lijffjit, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' "Dispersion and adjusted frequencies in corpora". *International Journal of Corpus Linguistics* 17(1). 147–149.

Lo Cascio, Vincenzo. 2012. *Dizionario Combinatorio Compatto Italiano*. Amsterdam: John Benjamins.

Lo Cascio, Vincenzo. 2013. *Dizionario Combinatorio Italiano*. Amsterdam: John Benjamins.

Mauri, Caterina, Silvia Ballarè, Eugenio Goria, Massimo Cerruti & Francesco Suriano. 2019. KIParla corpus: A new resource for spoken Italian. In Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds.), *Proceedings of the 6th Italian conference on computational linguistics (CLiC-it)*, vol. 2481. Bari, Italy: CEUR Workshop Proceedings.

Naismith, Ben & Alan Juffs. 2025. The impact of collocational proficiency features on expert ratings of L2 English learners' writing. *Studies in Second Language Acquisition* 47. 336–360.

Orenha-Ottaiano, Ana, Maria Garcia-Gonzalez, Maria Eugênia Olímpio, Marie-Claude L'Homme, Margarita Alonso Ramos, Cláudio R. Valencio & Walkiria Tenorio. 2021. Corpus-based methodology for an online multilingual collocations dictionary: First steps. In Iztok Kosem, Miloš Cukr,

Marek Jakubíček, Jelena Kallas, Simon Krek & Carole Tiberius (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual*, 1–28. Brno: eLex.

Paulsen, Mikel Ekeland. 2023. Assessing word commonness: Adding dispersion to frequency. *International Journal of Corpus Linguistics* 28(3). 318–343.

Pecina, Pavel. 2005. An extensive empirical study of collocation extraction methods. In Chris Callison-Burch & Stephen Wan (eds.), *Proceedings of the ACL student research workshop*, 13–18. Ann Arbor, Michigan: Association for Computational Linguistics.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources & Evaluation* 44(1). 137–158.

Perri, Damiano, Irene Fioravanti, Osvaldo Gervasi & Stefania Spina. 2024. Combining grammatical and relational approaches: A hybrid method for the identification of candidate collocations from corpora. In Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre & Alexandre Rademaker (eds.), *Proceedings of the joint workshop on multiword expressions and universal dependencies (MWE-UD) @ LREC-COLING 2024*, 138–146. Torino, Italia: ELRA and ICCL.

Qader, Wisam A., Musa M. Ameen & Bilal I. Ahmed. 2019. An overview of bag of words: Importance, implementation, applications, and challenges. In *Proceedings of the 2019 international Engineering Conference (IEC)*, 200–204. Erbil, Iraq: IEEE.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*, 101–108. Online: Association for Computational Linguistics.

R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rayson, Paul, Damon Berridge & Brian Francis. 2004. Extending the cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th international conference on statistical analysis of textual data*, 926–936. Louvain: Presses Universitaires de Louvain.

Römer Barron, Ute & Rainer Schulze (eds.). 2009. *Exploring the lexis-grammar interface*. Amsterdam: John Benjamins Publishing.

Rychlý, Pavel. 2008. A lexicographer-friendly association score. In Petr Sojka & Aleš Horák (eds.). *Raslan 2008: Recent advances in slavonic natural language processing. Proceedings of the second workshop on RASLAN, Karlova Studánka, Czech Republic, December 5–7, 2008*, 6–9. Brno: Masaryk University.

Salvi, Giampaolo. 2013. *Le parti del discorso*. Roma: Carocci.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*. Manchester, UK.

Seretan, Violeta. 2011. *Syntax-based collocation extraction. Text, speech and language technology 44*. Dordrecht: Springer.

Shi, Tianze & Lillian Lee. 2020. Extracting headless MWEs from dependency parse trees: Parsing, tagging, and joint modeling approaches. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics*, 8780–8794. Online: Association for Computational Linguistics.

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2017. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Proceedings of the 13th workshop on Multiword Expressions (MWE 2017)*, 48–53. Valencia, Spain: Association for Computational Linguistics.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, Anna. 2015. Collocation in beginner learner writing: A longitudinal study. *System* 53. 148–160.

Spina, Stefania. 2014. Il Perugia Corpus: Una risorsa di riferimento per l'italiano: Composizione, annotazione e valutazione. In Roberto Basili, Alessandro Lenci & Bernardo Magnini (eds.), *Proceedings of the first Italian conference on computational linguistics CLiC-it 2014*, vol. 1, 354–359. Pisa: Pisa University Press.

Spina, Stefania. 2016. Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. In Beatriz Sanromán Vilas (ed.), *Collocations cross-linguistically: Corpora, dictionaries and language teaching*, 219–244. Helsinki: Société Néophilologique. (Mémoires de la Société Néophilologique de Helsinki).

Spina, Stefania, Fabio Zanda & Irene Fioravanti. 2025. From PEC to PEC24: A new reference corpus for Italian. *Italiano LinguaDue* 17(1). 745–768.

Su, Qi, Chen Gu & Pengyuan Liu. 2024. Association measures for collocation extraction: Automatic evaluation on a large-scale corpus. *International Journal of Corpus Linguistics* 29(1). 59–86.

Tiberii, Paola. 2018 [2012]. *Dizionario delle collocazioni: Le combinazioni delle parole in italiano*, 2nd edn. Bologna: Zanichelli Editore.

Urzì, Francesco. 2009. *Dizionario delle combinazioni Lessicali*. Lussemburgo: Convivium.

Venables, William N. & Brian D. Ripley. 2002. *Modern applied statistics with S.*, 4th edn. New York: Springer.

Wahl, Alexander & Stefan Th. Gries. 2018. Multi-word expressions: A novel computational approach to their bottom-up statistical extraction. In Pascual Cantos-Gómez & Moisés Almela-Sánchez (eds.), *Lexical collocation analysis*, 85–108. Cham: Springer.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zanda, Fabio. 2025. *Assessing receptive collocation knowledge in learner Italian: Developing corpus-based tests for use with intermediate-advanced learners*. Perugia, Italy: University for Foreigners of Perugia PhD dissertation.