DESIGNING A CORPUS-BASED SYLLABUS OF ITALIAN COLLOCATIONS: CRITERIA, METHODS AND PROCEDURES[1]

FRANCESCA LA RUSSA[2], VERONICA D'ALESIO[3], ANNA SUADONI[4]

Abstract. Lexical combinations are central to language learning because they can be processed quickly (Siyanova-Chanturia 2013) and their use gives the idea of fluency in production (Nattinger and DeCarrico 1992). However, the acquisition of L2 phraseological competence is difficult for learners. This is particularly true for collocations, "sequences of words that tend to occur in stable and privileged combinations" (Simone 1990: 440). The difficulty is partially due to the fact that, unlike other types of lexical combinations, such as idiomatic phrases and proverbs, collocations are usually not emphasized in language courses, so learners do not notice them and do not assimilate them as complex lexemes (Bini *et al.* 2007). To fill this gap and provide teachers with a useful reference point for teaching collocations in language courses, we designed a corpus-based syllabus of Italian collocations. In this article, we illustrate the methodological choices and criteria followed for the realization of the syllabus.

Keywords: collocation, Italian, syllabus.

1. INTRODUCTION

Since Pawley and Syder's (1983) pioneering work on *sentence stems*, the interest in phraseological units has been increasingly growing. Soon after, two major shifts that had a huge impact on second language acquisition research occurred: on the one hand, the use of corpora showed that words systematically group together in a given language (Sinclair 1991) and that corpora themselves could be used as a powerful tool in language learning (Johns 1991; 2002); on the other hand, phraseological units gained momentum in teaching theories, most notably in Nattinger and DeCarrico's crucial work on lexical phrases (1992) and in Lewis' lexical approach (1993).

The term *phraseological units* refers to a variety of multiword expressions, ranging from idioms and fixed strings (e.g. over the moon, meet cute) to linguistic routines and

[1] The present work stems from a close cooperation between the authors. For the specific concerns of the Italian Academy, Francesca La Russa is responsible for paragraph §3; Veronica D'Alesio is responsible for paragraphs §1 and § 4; Anna Suadoni is responsible for paragraph § 2.

[2] Sapienza - Università di Roma, francesca.larussa@uniroma1.it.

[3] Sapienza - Università di Roma, veronica.dalesio@uniroma1.it.

[4] Universidad de Granada, asuadoni@go.ugr.es.

collocations (e.g. pay attention, highly recommend, wide awake). These combinations are characterized by various degrees of frequency, fixedness, and strength of association between the words they are made of and these features are known to affect both language learning and processing. A large body of psycholinguistic evidence shows that words that are frequently paired together and that are strongly associated with one another are processed more easily both in L1 and in L2 (Siyanova-Chanturia 2013). Moreover, as phraseological units stem from a native-like intuition (Meunier 2012), they are an essential part of the production and perception of fluency in a second language.

Given the undeniable utility of phraseological units in speech and language learning, this article presents a syllabus of Italian verb-noun collocations. Unlike fixed phraseological sequences, most of the collocations allow for some freedom in the choice of collocates, while they still maintain a certain degree of semantic transparency. In this sense, collocations can be considered as prototypical schemata that can be learned and that can later turn into fully productive schematic patterns, rather than be memorized as frozen units. As stated by Ellis (2012) a distinction can be made between *targets for learning* and *seeds of learning*, where the former – like nontransparent less frequent idioms – are hard to master even for L2 advanced speakers, and the latter – more conventionalized prototypical sequences – are readily learnable and feel safe to use thanks to their high frequency in the target language. Therefore, collocations, being easier to memorize and reuse, can be seen as a more approachable way to get access to native-like formulas.

In this article, at first, a definition of collocations will be given and the advantages as well as the challenges of teaching them in second language classes will be discussed. Our proposal for a corpus-based syllabus of Italian collocations will be later illustrated in detail. After briefly presenting the possible models, the methodological choices will be discussed. In particular, in compiling the syllabus, both CELI – a learner corpus of intermediate and advanced second language speakers – and PEC – a larger oral and written corpus of native speakers' productions – were chosen as reference corpora. The reason is twofold: by checking learners' productions against a native corpus, we ensure that the use of those items is actually well represented in the language, while simultaneously providing reliable data on the proficiency level of emergence of a given collocation. Finally, the questions that arose while compiling the syllabus, along with the implications for future research will be discussed.

2. DEFINING, TEACHING AND LEARNING COLLOCATIONS

According to the idiom principle (Sinclair 1991: 110), in communicative events, the speaker is naturally subject to a series of combinatorial restrictions that give rise to partially prefabricated structures belonging to different lexical categories: "The principle of idiom is that a language user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments."

Discourse is therefore formed by a chain of words which, at the syntactic-semantic level, tend to occur in a limited number of combinations with other words, creating sequences with a varying degree of cohesion.

Obviously, for reasons of coherence and semantic proximity, the context may limit the lexical options available to the speaker. However, in many cases the frequency with which words combine with other words cannot be explained only by the relationship with the concepts to be described and represented in the communicative event.

Lexical units formed by combinations of words with a high degree of fixation and whose co-occurrence cannot be predicted on the basis of semantic criteria only are frequent in texts. Their definitions and categorizations are numerous and heterogeneous. In this study, the term *phraseological units* will be adopted.

Simone (1990) identifies three different types of phraseological units:

- complex or multi-word units: sequences made of two or more words that syntactically function as a single word;
- collocations: very frequent sequences of words that tend to occur in stable and privileged combinations;
- idioms: sequences of words in which the meaning of the whole cannot be deduced compositionally from the meaning of the single elements.

Degree of fixation, compositionality of the meaning and idiomaticity are often adopted as criteria of distinction and definition of phraseological units. However, classifying them into discrete units is impossible. We would rather speak of a continuum whose endpoints are limited on the one side by free combinations and on the other by idiomatic expressions and multi-word units. The latter cannot be modified either at the syntagmatic or at the paradigmatic level and their meaning is global and not transparent.

Between these two extremities, there are preferential combinations, closer to free combinations, and collocations. Collocations are subject to greater combinatorial restrictions at the paradigmatic level, but can be modified at the syntagmatic level (passivization, dislocation, interposition of other elements, etc.) (Simone 2007; Masini 2009). Concerning the semantic level, they are rather transparent given the compositionality of the meaning.

One of the two elements that make up the collocation is autonomous while the other fully realizes its meaning in combination only (Hausmann 1979). According to Mel'čuk (1998: 29):

> A collocation AB of L is a semantic phraseme of L such that its signified 'X' is constructed out of the signified of the one of its two constituent lexemes — say, of A — and a signified 'C' [X = A+C] such that the lexeme B expresses 'C' contingent on A [5].

Linguistic use entails the restriction and results in the privileged but non-exclusive combination that characterizes collocations. This process increases the difficulty of establishing discrete limits that clearly separate the collocations from the two extremities of the continuum.

In the case of collocations, some words tend to bind themselves to other words without the need for any obligatory co-presence in the syntagma or semantic implications between the two elements. Furthermore, collocations generally lack idiomaticity since the elements that compose them maintain a certain semantic independence.

The rigidity of the restrictions that concern collocations is halfway between idioms and free combinations. Actually, as Martin (1992: 157) explains, the inclusion of collocations in the category of phraseological units is not unanimous: "They [collocations] neither can be considered to be idioms, nor that they can be regarded as free word groups. Typically they are in-between: restricted enough not to be regarded as free, transparent enough not to be considered idiomatic."

[5] A and B are lexical items, L denotes a language.

Collocations can be classified on the basis of the elements they are made of:

- Verb + noun (complement): *chiedere aiuto* 'ask for help';
- Noun (subject) + verb: *la guerra scoppia* 'the war breaks out';
- Noun + adjective: *errore madornale* 'huge mistake';
- Adjective + noun: *alta opinione* 'high opinion';
- Noun + noun: *parola chiave* 'keyword';
- Noun + preposition + noun: *stormo di uccelli* 'flock of birds';
- Adverb + adjective: *fermamente convinto* 'firmly convinced';
- Verb + adverb: *pentirsi amaramente* 'bitterly regret'.

Verb + noun collocations – that are the object of our syllabus – are the first of the four main types of collocations according to Mel'čuk's classification (1998: 29): when the signified C assumed by B in a collocation AB is different from the signified given by B in the dictionary, there are two possibilities:

> […] a. 'C' is empty, that is, the lexeme B is, so to speak, a semi-auxiliary selected by A to support it in a particular syntactic configuration; or b. 'C' is not empty but the lexeme B expresses (C) only in combination with A (or with a few other similar lexemes)].

When the meaning of the collocated verbs is 'empty', according to Mel'čuk (1998), they can be defined as light (D'Agostino 1993; Cicalese 1999). Light-verbs are highly frequent and polysemous, and have a basic meaning. In some cases, they don't provide any information at a semantic level, just expressing the grammatical marks for the action expressed by the noun (e.g. *fare una passeggiata* 'take a walk' = *passeggiare* 'stroll'); in other cases, they can be combined with predicative nouns that are not related to verbs at the morpho-phonological level (*fare la doccia* 'have a shower' = ?). The most frequent light verbs in Italian are: *essere* 'be'; *fare* 'do'; *prendere* 'take'; *dare* 'give'; *mettere* 'put'; *portare* 'bring'; *avere* 'have'. Other verbs, the so-called extensions of the light-verb (Cicalese 1999), can assume the same semantic-syntactic function in certain combinations. These verbs are generally used with a predicative function that, in some cases, can replace a light verb. Although at the syntactic level the role of light-verbs and light-verb extensions are equivalent, the latter are often used in more formal contexts (*fare un accordo* 'make a deal' vs. *concludere un accordo* 'close a deal').

Managing these constructions – that are tightly related to language use and to the pragmatic aspects of communication – is natural for native speakers (Coseriu 1977; Nattinger and DeCarrico 1992; Wray 2002; Schmitt 2004), while it is much more difficult for non-native speakers.

The combinatorial restrictions that operate in collocations – although common among languages – might be an obstacle to their acquisition by non-native speakers. As a matter of fact, the relationship between the elements that make up the collocation is institutionalized by linguistic use and, as a consequence, not predictable. As Jezek (2016: 193) explains:

> They are restricted by a constraint that seems to be rooted in language use, that is, in the tendency of languages to express a given content by means of preferential word pairs, although other combinations are in principle possible from a semantic perspective.

However, their communicative potential makes their inclusion in a foreign language curriculum indispensable for the development of an adequate socio-pragmatic competence and for improving learners' fluency.

The need of addressing the study of vocabulary in a combinatorial perspective, based on more complex units of analysis, is supported by several studies that suggest that words are stored and organized by learners as part of a network connected by different types of relationships (Aitchison 1987). Knowing a word also means knowing its collocational meaning (Nation 1990; 2001). Drawing learners' attention to the possibility or impossibility of certain combinations would therefore help them construct their own lexical repertoire, avoiding interferences with the mother tongue and increasing the probability of long-term memorization. Hill (2000: 61-62) offers the example of the verb "speak":

As we saw above, with a common verb like "speak" we cannot say that students really know the word unless they know at least the following possibilities:

- Speak a foreign language
- Speak (French)
- Speak fluently
- Speak your mind
- Speak clearly
- Speak with a (Welsh) accent
- Speak in public
- Speak openly
- Speak volumes

3. DESIGNING A SYLLABUS OF ITALIAN COLLOCATIONS

A syllabus of Italian collocations would provide teachers with a useful reference for teaching collocations in language courses, allowing them to draw the students' attention to a pervasive linguistic phenomenon.

The term syllabus refers to the specification and sequencing of teaching contents in terms of knowledge and/or skills (Ciliberti 2012). More specifically, in language teaching it indicates the list of the linguistic-communicative elements that students have to acquire at a given proficiency level.

Going through some of the main Italian L2 syllabi and profiles, it is easy to notice that collocations do not have much space. For example, the *Profilo della lingua italiana*[6] (Spinelli and Parizzi 2010) only provides lexical lists of single words in alphabetical order for levels from A1 to B2, while in the *Sillabo di riferimento per i livelli di competenza in Italiano L2* (AA. VV., 2011), a non-exhaustive list of words (for example, *padre* 'father'; *madre* 'mother'; fratello 'brother', etc.) is placed alongside each semantic area (for example, *famiglia*, 'family').

[6] The *Profilo della lingua italiana* (Spinelli and Parizzi 2010) is the official Italian Reference Level Description (https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions).

Even though it is not a proper syllabus, higher importance is given to collocations in the *Dizionario delle collocazioni italiane per apprendenti* (DICI-A, cf. Spina 2016). The DICI-A is a corpus-based and specifically targeted to learners dictionary of Italian collocations. The collocations included in the DICI-A were extracted from the *Perugia corpus* (PEC, cf. Spina 2014), a reference corpus that includes written and oral texts from different textual genres. The collocations were then ordered by their coefficient of usage that combines the measure of frequency with the measure of dispersion through the different textual genres in the corpus (for more information cf. Spina 2016). Finally, they were assigned to a given proficiency level based on their coefficient of usage and their topic and function. The initial goal was to assign lexical combinations to three proficiency levels: level A (Basic User), level B (Independent User) and level C (Proficient User). The only completed level to date is level A. The combinations assigned to this level are:

- made up of words that belong to level A of the *Profilo della lingua italiana* (Spinelli and Parizzi 2010);
- positioned at the highest ranks of the coefficient of usage;
- related to topics of the most immediate relevance (e.g. basic personal and family information, shopping, local geography, employment, etc.).

With its solid methodology that combines empirical and intuitive processes, the DICI-A could be a valuable model for the creation of our syllabus of Italian collocations.

The syllabus is based both on learners and native speakers corpora and has been designed relying upon both empirical data and qualitative judgments. In the following paragraphs, the methodological criteria adopted for the selection and sequencing of its contents will be discussed in further detail.

3.1. Content selection

The first decision to make concerns the choice of the reference corpus for the extraction of the collocations. Two main options exist:

- extracting collocations from a learner corpus, thus including the collocations that are actually produced by learners at a given proficiency level;
- extracting collocations from a corpus that collects native speakers' productions, thus indicating which collocations learners should use at a given proficiency level based on their frequency in the native speakers' corpus.

Extracting collocations from a learner corpus provides reliable data on learners' authentic use of the language and shows direct evidence of when collocations begin to be used by learners becoming part of their productive repertoire. Furthermore, as stated by Capel (2010: 4): "An analysis based solely on native speaker frequency does not capture certain words that are useful to learners and which have a high frequency in the language classroom." Therefore, following the model of the *English Vocabulary Profile* (*EVP*)[7], collocations were extracted from a learner corpus.

Nonetheless, the frequency of occurrence of a given collocation in native speakers' production – and consequently in the input learners are exposed to – plays a crucial role in

[7] An online interactive resource (https://www.englishprofile.org/wordlists) describing the vocabulary used by learners of English at each proficiency level. Since, as affirmed by the authors " its aim is to reflect what learners do know, not what they should know",the EVP is primarily based on the *Cambridge Learner Corpus* (*CLC*), a collection of several hundred thousand examinations attended by students of English from all over the world and on other sources related to English as a second language, such as classroom materials.

its acquisition (Ellis 2002). Usage-based theories on language acquisition claim that more frequent lexical elements lead to stronger mental representations. When it comes to L2 learners, it is acknowledged that they acquire phraseology through exposure and repetition. As Spina (2020: 41-42) explains: "the more learners are exposed to a phraseological sequence, the stronger it becomes entrenched in their memory and the easier it is accessed, processed and produced." Consequently, the exposure to word combinations that are frequent in the input is likely to impact their production by learners. As a matter of fact, the knowledge of phraseological sequences is strongly correlated with frequency values in native corpora (Durrant 2014) and experimental studies (e.g. Spina 2015) show that learners tend to overuse frequent combinations to which they are exposed many times. For these reasons, frequency in native speakers' productions was also taken into account when assigning a collocation to a given proficiency level. We will discuss this point in further detail in §3.2.

The learner corpus from which collocations were extracted is the CELI corpus (Spina *et al.* 2022), a balanced pseudo-longitudinal corpus that collects 3041 written texts produced by learners of Italian L2 who attended the *Certificati di Lingua Italiana* (CELI) exams (levels B1, B2, C1, C2). The main corpus is made up of four sub-corpora. Each sub-corpus collects the written productions corresponding to a given level. The automatic extraction from the corpus involved three Part Of Speech (POS) sequences:

- noun-adjective: *sistema operativo* 'operating system';
- verb-adverb: *tornare indietro* 'go back';
- verb-noun: *prendere una decisione* 'make a decision'.

At the moment, only the verb-noun collocations have been filtered for inclusion in the syllabus. The list of the automatically extracted verb-noun combinations included collocations (*prestare attenzione* 'pay attention'), some light-verb constructions (*fare colazione* 'have breakfast') and idioms (*abbandonare la nave* 'jump ship') that were kept in the syllabus. However, it also included some non-target-like combinations (e.g. **utilizzare attenzione* '*use attention', instead of *prestare attenzione* 'pay attention') and free combinations (e.g. *cercare televisione* 'look for television') that had to be removed.

The raw list was thus further filtered combining both objective statistical measures of association and frequency and intuitive phraseology judgments. The measure that was adopted to eliminate most of the free word combinations was that of strength of association, i.e. "the degree to which two words are tightly and exclusively connected to one another" (Spina 2020: 43). Strength of association is usually operationalized in the *Pointwise Mutual Information* (PMI) score, a measure that "compares the probability of observing word *a* and word *b* together with the probabilities of observing *a* and *b* independently" (Paquot 2019: 5). It is calculated by comparing the observed number of occurrences of a word pair with its expected number of occurrences[8] and is generally used as a measure of collocational strength since it brings out word combinations made up of closely associated words. According to this approach, a PMI score of 3 or above indicates a significant collocation threshold (Hunston 2002; Stubbs 1995). For this reason, all the collocations with a PMI value below 3 were removed from the initial list of verb-noun collocations.

To eliminate most of the non-target-like collocations, the coefficient of usage of the collocations in the *Perugia corpus* (Spina 2014) was considered. As previously explained,

[8] It is calculated as follows: PMI = log2 P(w1|w2) / P(w1)P(w2). Where Log2 stands for logarithm base 2, P for probability, w1 for word 1 and w2 for word 2.

this measure accounts for the frequency and dispersion of the collocations in native speakers' written and oral productions. Following the model of the DICI-A (Spina 2016), all the collocations with a coefficient of usage below 2 were removed from the list.

After this first screening, five linguists were asked to judge the remaining combinations. Based on the idea that conventional combinations are often extremely useful for L2 learners, it was decided to keep in the final list not only pure collocations but also some word combinations that are highly conventional, such as *aprire la porta*, 'open the door'; *chiudere la porta,* 'close the door'; *lavare i piatti*, 'wash the dishes' and so on.

After this further screening, the final syllabus list is made up of 953 collocations, highly conventional combinations, light-verb constructions and some idioms.

3.2. Compiling methods

In the following paragraphs, the criteria and the procedure adopted to pinpoint the proficiency level at which collocations should be taught/learned as well as the final organization of the syllabus will be described.

3.2.1. Compiling criteria

The collocations in the final list had to be assigned to the CEFR levels from B1 to C2. In order to do so, an integrated approach that relies upon both empirical data and qualitative judgments was adopted and several criteria were followed.

The two main criteria are the frequency of the collocation in native speakers' productions, and the proficiency level at which it is used more often by learners. Regarding the first criterion, to account for the frequency of the collocations in native speakers' productions, their coefficient of usage in the *Perugia corpus* (Spina 2014) was considered and three bands – corresponding to low, mid and high frequent collocations – were created. For example, the collocation *trovare lavoro* ('find a job', coefficient of usage: 111) belongs to the high frequency band; *visitare città* ('visit a city', coefficient of usage: 6) belongs to the mid frequency band and *sottovalutare importanza* ('underestimate the importance', coefficient of usage: 2) belongs to the low frequency band.

The second criterion is based on learners' productions and concerns the analysis of the collocations that were actually produced by learners at different proficiency levels. Since a given collocation is often used at different proficiency levels, the number of occurrences of the collocation in the four CELI subcorpora corresponding to levels B1, B2, C1 and C2 was observed and the level with more occurrences was pinpointed. For example, the collocation *trovare lavoro* ('find a job') is used 58 times at level B1, 39 times at level B2, 22 times at level C1 and 9 times at level C2. Thus, the pinpointed level is B1.

Since the first and second criterion might give contrasting information, for example when a very frequent collocation in the native corpus is more often used at a high proficiency level (e.g. C1 or C2) or vice versa, two further criteria were adopted: the presence of the words that make up the collocation in the lexical lists of the *Profilo della lingua Italiana* (Spinelli and Parizzi 2010) and the topic addressed by the collocation.

The lexical lists of the A1, A2, B1 and B2 levels of the *Profilo della lingua Italiana* (Spinelli and Parizzi 2010) – that is the official Italian Reference Level Descriptions – were chosen as a reference. Notably, it was observed to which proficiency level the words that make up the collocation belong. In the case of the collocation *trovare lavoro* ('find a job'), for example, both *trovare* and *lavoro* belong to the lexical lists of level A1.

This criterion was more useful with transparent combinations (e.g. *trovare lavoro*, 'find a job') then with more opaque ones (such as *fare strada*, 'lead the way'). As a matter of fact, if beginner learners know the words *fare* (literally *'make*') and *strada* ('street'), this does not mean that they also know the collocation *fare strada*. Nevertheless, it can be assumed that the collocation is composed of words that learners already know.

Finally, the topic to which each collocation refers was described and taken into account when pinpointing the proficiency level. In particular, based on the information on vocabulary range and control provided by the CEFR, it was cross-checked that the collocations that were assigned to a given proficiency level refer to topics that are relevant for that level. For example, the collocations assigned to level B1 should address topics relating to "[…] everyday life such as family, hobbies and interests, work, travel, and current events" (Council of Europe 2020: 132). The topic description was made following the classification made by the Italian Profile which provides a list of the words that express a given notion at different proficiency levels (e.g. for the notion *Tempo libero e intrattenimenti*, 'Free time and entertainment', it indicates *uscire*, 'go out' at level A1; *gita*, 'trip' at level A2; *serata*, 'event' at level B1 and so on). Once completed, the syllabus of collocations will allow the phraseological dimension to be integrated to the already available lists of words of the Profile.

For example, the collocation *trovare lavoro* ('find a job') refers to the topic *Ricerca di un posto di lavoro* ('Job seeking') that is relevant for B1 learners.

In conclusion, research based on the CELI corpus, coupled with the research into native speaker frequency, the Italian Reference Level Descriptions and topic analysis allowed to pinpoint the most appropriate proficiency level for the collocations.

3.2.2. Compiling procedure

The procedure that was adopted to assign each collocation to a given proficiency level might be synthesized as shown in Table 1.

Table 1.

Procedure adopted to assign collocations to a proficiency level

Coefficient of usage in native speakers' production	Check if the collocation belongs to the high, medium or low frequency band: ● collocations in the high frequency band should be assigned to level B1 or level B2; ● collocations in the medium frequency band should be assigned to level B2 or level C1; ● collocations in the low frequency band should be assigned to level C1 or C2.
Number of occurrences in the CELI subcorpora	Between the two proficiency levels indicated by the frequency band, assign the collocation to the level in which it occurs more often. If the number of occurrences in the two levels is the same or similar (1 or 2 occurrences of difference) or the collocation is already used consistently (e.g. 10 times) at the lower level, assign it to the lower level.

Italian Profile	When criteria 1 and 2 give contrasting information, check the Italian Profile lexical lists and assign the collocation to the level to which the words that make up the collocation belong.
Topic	Double check if the collocations assigned to a given proficiency level address topics that are relevant to that level.

The process of connecting the word combinations to proficiency levels can be exemplified through the following cases:

- *trovare lavoro* 'find a job': belongs to the high frequency band. Therefore it should be assigned to level B1 or B2. It is used 58 times at level B1 and 39 times at level B2. The addressed topic (job seeking) is relevant for B1 learners, thus it was assigned to level B1;
- *avere diritto* 'have right to': belongs to the high frequency band. As a consequence, it should be assigned to level B1 or B2 but it is never used at level B1. It is used 5 times at level B2, 40 times at level C1 and 20 times at level C2. The word *diritto* does not appear in the lexical lists of the Italian Profile, we can therefore assume that it is learned at an advanced level. The topic addressed is that of socio-political structures, thus it is more relevant for C level learners. Since the collocation is used more often at level C1, it was assigned to this level;
- *visitare città* 'visit a city': belongs to the medium frequency band, consequently, it should be assigned to level B2 or C1. It is used 12 times at level B1, 6 times at level B2 and 7 times at level C1. Both *visitare* and *città* belong to the A1 lexical list of the Italian Profile and the topic relates to travel and everyday life. Since the collocation is already consistently used at level B1, B1 learners already know the words that make up the collocation and the topic is relevant for this level, the collocation was assigned to level B1.

3.3. Organization of the syllabus

The final result is a syllabus in which the 953 collocations are organized according to both proficiency level and topic. 221 collocations were assigned to level B1, 369 to level B2, 299 to level C1 and 63 to level C2.

Furthermore, the 953 collocations were distributed among 70 topics, so that it is also possible to search for all the collocations related to a specific topic. For example, the collocations belonging to different proficiency levels that were assigned to the topic *Spesa – Prezzi e strumenti di pagamento* 'Expenses – Prices and payment instruments' are:

- *at* level B1: *pagare conto* 'pay the bill'; *pagare prezzo* 'pay the price'; *pagare affitto* 'pay the rent'.
- at level B2: *abbassare prezzo* 'lower the price'; *fare soldo* 'make money'; *mantenere famiglia* 'feed a family'; *pagare bolletta* 'pay the bills'; *pagare tassa* 'pay taxes'; *risparmiare soldo* 'save money'; *spendere soldo* 'spend money'.
- at level C1: *buttare soldo* 'waste money'.

4. CONCLUSION

Phraseological units are a large part of language use and have been the focus of second language acquisition research for many years. Despite being a key factor in learning new vocabulary and gaining fluency, multiword units are still scantily represented in syllabi and course materials for Italian as a second language.

To fill this gap, we presented a syllabus of Italian collocations for intermediate and advanced learners, targeting four CEFR levels (from B1 to C2). We adopted a descriptive rather than a prescriptive approach, drawing data from CELI (Spina *et al.* 2022), a pseudo-longitudinal learner corpus of written texts arranged in four sub-corpora, one for each CEFR level here considered. An initial list of verb-noun(obj) combinations was automatically extracted and then screened according to the *Pointwise Mutual Information* score (cutoff ≥3, cf. Stubbs 1995; Hunston 2002) and coefficient of usage in PEC (Spina 2014), a native reference corpus (cutoff ≥ 2).

Several studies (Ellis 2002; Durrant 2014; Spina 2020) highlight the importance of exposure in acquiring new vocabulary: particularly usage-based theories of language acquisition assume that the more frequently a form is encountered, the more entrenched in memory it will be. Following this perspective, in assigning each collocation to one of the four CEFR levels considered in this study, its frequency of occurrence in the native corpus was given a key role. Four major criteria were chosen to pinpoint the CEFR level of a given collocation:

- its coefficient of usage in the native corpus (whether it belongs to a high, medium, or low frequency band);
- the number of occurrences in each of the four learner sub-corpora;
- whether the single words that make up the collocation are present in the lexical lists of the Italian Profile (Spinelli and Parizzi 2010) and at which CEFR level;
- whether the collocation falls within a topic that is relevant to that CEFR level.

The syllabus consists of 953 collocations: 221 collocations assigned to level B1, 369 to level B2, 299 to level C1 and 63 to level C2. Furthermore, collocations can be sorted by their proficiency level and by topic(s), in line with integrated approaches to syllabus design that highly value the communicative and the socio-pragmatic dimensions.

Most certainly, some limitations need to be considered. As stated earlier, the collocations were selected based on statistical measures and a final screening made by experts: while individual ratings inevitably leave room for subjectivity, when dealing with language – especially if targeted to learners – empirical data can be used as a guide up to a certain point. Focusing on the scope of the syllabus, an array of combinations that can actually be useful in the classroom and in the everyday life of Italian learners at various proficiency levels were included. This resulted in a list made up of collocations, as well as conventional combinations, idioms and light-verb constructions, making it clear that conventionality and semantic transparency balance out on a gradient scale.

Finally, the work presented in this paper will hopefully be a starting point for future developments: in particular, the collection of larger learner corpora for Italian as a second language could allow for further research on lower proficiency levels (corresponding to the A CEFR levels). Moreover, we hope that the methodology presented in this article could be a baseline for the study of other types of phraseological combinations (e.g. noun-adjective

collocations). As there is general consensus on the pragmatic value of formulaic sequences in language acquisition, we believe that the use of corpora in second language research could help provide reliable tools for researchers, learners and teachers alike.

ACKNOWLEDGMENTS. The research leading to these results received funding from PRIN-PHRAME Project (20178XXKFY) Phraseological Complexity Measures in learner Italian – Integrating eye-tracking, computational, and learner corpus methods to develop second language pedagogical resources.

REFERENCES

Aitchison, J., 1987*, Words in the Mind: An introduction to Mental Lexicon*, Oxford, Basil Blackwell.

Capel, A., 2010, "A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project", *English Profile Journal,* 1, 3, 1–11. https://doi.org/10.1017/S2041536210000048

Cicalese, A., 1999, "Le estensioni di verbo supporto. Uno studio introduttivo", *Studi italiani di linguistica teorica ed applicata*, 28, 3, 447–485.

Ciliberti, A., 2012, *Glottodidattica. Per una cultura dell'insegnamento linguistico*, Firenze, Carocci Editore.

Coseriu, E., 1977, *Principios de semántica estructural*, Madrid, Gredos.

Council of Europe, 2020*, Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors,* Strasbourg, Council of Europe Publishing.

D'Agostino, E., 1992, *Analisi del discorso*, Naples, Loffredo.

Durrant, P., 2014, "Corpus frequency and second language learners' knowledge of collocations : a meta-analysis", *International Journal of Corpus Linguistic*, 19, 4, 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

Ellis, N., 2002, "Frequency effects in language processing and acquisition", *Studies in Second Language Acquisition,* 24, 2, 143–188. https://doi:10.1017/S0272263102002024

Ellis, N., 2012, "Formulaic language and second language acquisition: Zipf and the phrasal teddy bear", *Annual review of applied linguistics*, 32, 17–44. https://doi:10.1017/S0267190512000025

Hausmann, F. J., 1979, "Le dictionnaire de collocations", in F.J. Hausmann et al. (eds.). *Wönterbücher. Dictionaries. Dictionnaires. Ein internationals Handbuch zur Lexikographie. Enciclopédie international de lexicographie, vol. I,* Berlin-New York, Walter de Gruyter, 1010–1019.

Hill, J., 2000, "Revising priorities: from grammatical failure to collocational success", in: M. Lewis (ed.). *Teaching collocation. Further Developments in the Lexical Approach*, London, Language Teaching Publications, 47–69.

Hunston, S., 2022, *Corpora in applied linguistics*, Cambridge, Cambridge University Press.

Ježek, E., 2016, T*he lexicon: An introduction*, Oxford, Oxford University Press.

Johns, T., 1991, "Should You Be Persuaded: Two Examples of Data-Driven Learning Materials", *English Language Research Journal*, 4, 1–16.

Johns, T., 2002, "Data-driven Learning: The Perpetual Challenge", in: B. Ketteman, G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis,* Leiden, The Netherlands, Brill, 105–117.

Lewis, M., 1993, *The lexical approach*, Hove, England, Language Teaching Publications.

Masini, F., 2009, "Combinazioni di parole e parole sintagmatiche", in: E. Lombardi Vallauri, L. Mereu (eds), *Spazi linguistici. Studi in onore di Raffaele Simone,*Roma, Bulzoni, 191–209.

Martin, W., 1992, "Remarks on Collocations in Sublanguages", *Terminologie et Traduction,* 2, 3, 157–164.

Mel'čuk, I., 1998, "Collocations and lexical functions", in: A. P. Cowie (ed.), *Phraseology. Theory, analysis, and applications*, Oxford, Clarendon Press, 23–53.

Meunier, F., 2012, "Formulaic language and language teaching", *Annual review of applied linguistics*, 32, 111–129. http://doi:10.1017/S0267190512000128

Nation, I. S. P., 1990, *Teaching and Learning Vocabulary*, Boston, Heinle & Heinle Publishers.

Nation, I. S. P., 2001, *Learning vocabulary in another language,* Cambridge, Cambridge University Press

Nattinger, J., J. DeCarrico, 1992, *Lexical phrases and language teaching*, Oxford, Oxford University Press.

Paquot, M., 2019, "The phraseological dimension in interlanguage complexity research", *Second language research*, 35, 1, 121–145. https://doi.org/10.1177/0267658317694221

Pawley, A., F. H. Syder, 1983, "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency", in: J.C. Richards, R.W. Schmidt (eds), *Language and Communication*, London, England, Routledge, 191–225.

Schmitt, N., 2004, *Formulaic sequences: Acquisition, processing and use*, Amsterdam, John Benjamins Publishing Company.

Simone, R., 1990, *Fondamenti di linguistica* (No. 9), Bari, Laterza.

Simone, R., 2007, "Categories and constructions in verbal and signed languages", in: E. Pizzuto, P. Pietrandrea, R. Simone (eds), *Verbal and signed languages. Comparing Structures, Constructs and Methodologies*, Berlin, Mouton de Gruyter, 197–252.

Sinclair, J., 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Siyanova-Chanturia, A., 2013, "Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings", *The Mental Lexicon*, 8, 2, 245–268. https://doi.org/10.1075/ml.8.2.06siy

Spina, S., 2014, "The dictionary of Italian collocations: Design and integration in an online learning environment", *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10), 3202–3208.

Spina, S., 2015, "Phraseology in academic L2 discourse: The use of multi-words units in a CMC university context", in: E. Castello, K. Ackerley, F. Coccetta (eds), *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*, Bern, Peterlang, 279–294.

Spina, S., 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: B. Sanromán Vilas (ed.), *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching* (Mémoires de la Société Néophilologique de Helsinki), Helsinki, Société Néophilologique, 219–244.

Spina, S., 2020, "The role of learner corpus research in the study of L2 phraseology: main contributions and future direction", *Rivista di Psicolinguistica Applicata*, 20, 2, 35–52. https://dx.doi.org/10.19272/202007702003

Spina, S., I. Fioravanti, L. Forti, V. Santucci, A. Scerra, F. Zanda, 2022, "Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2", *Italiano LinguaDue*, 14, 1, 116–138. https://doi.org/10.54103/2037-3597/18161

Spinelli, B., F. Parizzi, 2010, *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2*, Milano, La Nuova Italia.

Stubbs, M., 1995, "Collocations and semantic profiles: On the cause of the trouble with quantitative studies", *Functions of language,* 2, 1, 23–55. https://doi.org/10.1075/fol.2.1.03stu

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