cography tain ions Speech ologisms Idioms NLP Etymology Glossary Lemma Meaning gin Corpora Dictionary Word Lexicon ble Pronounciation Definition Headword amples Entry Lexicology Dictionary Use Lexical Resources

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

The interaction of argument structures and complex collocations: role and challenges in learner's lexicography

Giacomini L.[1,2]**, DiMuccio-Failla P., Lanzi E.**[2]

[1] *University of Hildesheim, Germany*
[2] *University of Heidelberg, Germany*

Abstract

This contribution focuses on the status of complex collocations in pattern-based learner's dictionaries, reporting on findings of the ongoing corpus-based project *Pattern-based learner's lexicography* (Hildesheim University/Heidelberg University). After comparing recursively built complex collocations with argument-related complex collocations, the paper concentrates on the latter type and its functions. On the one hand, complex collocations displaying argument complementarity efficiently support the identification and formulation of sense patterns. On the other hand, they can serve different purposes within the microstructure of a pattern-based dictionary, namely as semantic types of sense patterns or as lexicographic items in a subordinate treatment unit. Argument-related complex collocations are phraseological lexicalisations of the conceptual scenes provided by sense patterns, and are therefore of key importance to language learners. The challenges related to the extraction of complex collocations from corpora are also addressed in the paper, and proposals are made for improving time efficiency, coverage, and quality of extracted candidates in future research.

Keywords: learner's lexicography; sense pattern; complex collocation; argument structure; cognitive lexicography; lexicogrammar

1 Introduction

Studies on the treatment of collocations in lexicography have been relatively frequent in recent decades, partly following the constant development of new approaches and tools in corpus linguistics. However, the traditional view of collocations as simple, binary combinations still dominates contemporary paper and electronic dictionaries. In particular, the potential of complex collocations for learner's lexicography has remained largely unexplored in relevant literature. This paper deals with the topic of complex collocations from the specific angle of their interaction with argument structures of verbs, specifically aiming to illustrate their advantages for learner's lexicography. This topic is related to the ongoing corpus-based project *Pattern-based learner's lexicography* carried out at Hildesheim University and Heidelberg University, and aimed at the compilation of an electronic pattern-based learner's dictionary of Italian.

The primary theoretical background of the project is provided by the tradition of linguistic approaches covering the interplay between lexis and grammar (cf. among others, Halliday 1992; Gross 1994; Herbst 2016/2017; Herbst et al. 2014), as well as by cognitive lexicography (Geeraerts 2007; Ostermann 2015). At the core of the proposed microstructural model is the notion of *normal patterns of usage* (Sinclair 1996/2004; Hanks 2013), or *sense patterns*, as the true lexical units of language. A sense pattern (e.g. EN *to follow someone going somewhere*, DE *jdn./etw. mit den Augen verfolgen*, or IT *accompagnare qualcuno in un luogo/in un percorso*) is a syntactic-semantic entity given by a combination of syntactic and semantic arguments, semantic types, and semantic roles, uniquely identifies one meaning of a word, and has a largely phrasal nature (Sinclair 2004) [1]. In the pre-lexicographic stage of the project, the focus lies on exploring theoretical models which can help us cover the interplay between sense patterns and collocations in lexicography, on devising a semi-automated, corpus-based method for identifying and formulating sense patterns, as well as on designing the microstructure of verb entries. In performing these tasks, we take advantage of the contrastive analysis of verb patterns in Italian, German, English, and French. In this contribution, we will present data extracted from Italian and German corpora[2].

As pointed out in Giacomini & DiMuccio-Failla (2019), the method for sense pattern identification has changed over time, maximizing the use of corpora in the process of pattern-related data retrieval. In particular, the method changed from initial concordance analysis as proposed in the context of Corpus Pattern Analysis (CPA, Hanks 2004) and of CPA-oriented approaches (cf. Renau & Nazar 2016), to the analysis of collocates. This radical methodological change originates from the observation of patterns originally identified by manual analysis of concordances, a large amount of which appears to be of a phraseological nature. The use of collocations for detecting the correlation between argument structures and meanings of a verb has proven to be a reliable and highly accurate method, in which both simple and complex collocations play a key role. In doing this, we take into account corpus-based studies exploring possible associations between collocations and constructions: besides those directly inspired by Sinclair's *idiom principle* (cf. Stubbs 1995), also the ones carried out in the context of Pattern Grammar (cf. Hunston & Francis 2000), of Construction Grammar (cf. collostructional analysis, Stefanowitsch & Gries 2003, and the idea of language as a Collostructicon in

[1] A detailed discussion of the semantic and phraseological features of sense patterns is provided by DiMuccio-Failla & Giacomini (2017a/2017b).

[2] Examples will be followed by an English translation. For the sake of clarity, the abbreviations IT, DE, and EN for the three languages will be used throughout the paper.

Herbst 2018), as well as of Frame Semantics (cf. Almela-Sánchez 2019).

Section 2 of this paper discusses state-of-the-art approaches to complex collocations, focusing on their formation by recursive expansion or by argument complementarity. Section 3 concentrates on the advantages of complex collocations with argument complementarity for pattern-based learner's lexicography, discussing the challenges related to their extraction from corpora, as well as their possible metalexicographic and lexicographic applications. Section 4 provides some insights into initial tests for the evaluation of complex collocation coverage within the planned dictionary, and hints at future research developments aimed at the optimization of complex collocation extraction from corpora in the context of the lexicographic process.

2 The formation of complex collocations

The focus of our paper lies in the description of the metalexicographic and lexicographic use of complex collocations of verbs in a pattern-based learner's dictionary. In the context of lexicography- and computationally-oriented studies on collocations, complex collocations have generally been described or defined as collocations involving more than two content words, largely drawing on the traditional view of a collocation as a primarily binary word combination. While the general understanding of collocations as n-ary combinations has been gaining some popularity on the basis of corpus evidence in the recent past, the nature of complex, i.e. larger than binary, collocations appears not to have been explored in depth and to generally conform to a single description model.

2.1 Complex collocations built by recursive expansion

Complex collocations have been named in different ways, ranging from *grammatically extended collocations* (Tutin & Kreif 2016), *collocation chains* (Alonso Ramos & Wanner 2007), and *nested collocations* (Seretan 2011), up to *collocations of collocations* (Wehrli et al. 2010). Unfortunately, qualitative and quantitative lexicogrammatic approaches on which we rely for sense pattern identification (cf. Section 1) do not provide a specific focus on complex collocations. The topic of complex collocations sporadically appears in the context of very specific empirical studies, without being directly embedded in a distinct theoretical framework.

In his account of lexical combinatorics and challenges posed by the acquisition and application of collocational information, Heid (1994: 231) writes: "An additional problem of the interaction between syntactic and collocational description is the recursive nature of collocational properties: the components of a collocation can be again collocational themselves: next to the German collocation *Gültigkeit haben* (EN *to be applicable*) (n+v), we have *allgemeine Gültigkeit haben* (EN *to be generally applicable*), with the collocation *allgemeine Gültigkeit* (EN *general applicability*) (n+a) as a component. These cases have sometimes been analyzed as different from collocations, but there is no reason for such treatment."

The *recursive* nature of collocations implies that a core collocational phrase is progressively expanded by the addition of new collocates, like in the abovementioned example *allgemeine Gültigkeit haben* (Heid 1994: 231). As a result, simple collocations are embedded in complex ones (cf. also Seretan 2013). The typical case is the expansion by means of adjectival or adverbial modifiers. Simple collocations in this sense are potentially open to any extension, the only limit being the collocational range (McIntosh 1966) of the progressively added components. This natural limit will cause complex collocations to usually consist of a limited number of elements.

Heid (1994) also highlights the importance of a formal account of this kind of collocation, for instance for machine translation. In our opinion, the importance of complex collocations goes far beyond the issue of their formalization, and equally affects syntactic, semantic, and lexicological analysis. As later pointed out by Zinsmeister & Heid (2003) in the context of adjective-noun-verb triples, a frequent case is the combination of two collocations with the same base, e.g. in the German examples *eine klare Absage* + *Absage erteilen* (EN *a clear refusal* + *to give a refusal*), or *absolute Mehrheit* + *Mehrheit erreichen* (EN *absolute majority* + *to obtain a majority*). The same principle applies to other languages, for example to English (cf. *spark strong emotions* mentioned in Gouws 2015: 172).

Lexicography possibly provides the perfect environment for observing complex collocations, not only from a theoretical point of view but also from the empirical perspective of their extraction from corpora and their presentation to end users such as foreign language learners or translators.

2.2 Complex collocations built by argument complementarity

In the context of our lexicographic project, the application of the new pattern-based approach to meaning representation coincides with a new perspective on complex collocations. Our focus has been mainly on verbs as the substantial carrier of sense pattern structures, and it is exactly during our study of verb argument structures that we identified the significant role complex collocations can play in a learner's dictionary. In extracting, validating, and sorting collocations, we noticed that a different type of complex collocation can be identified and described for the purpose of verb pattern treatment. This type of complex collocation directly involves the level of argument structures: collocations extracted for at least two different arguments of a verb are often syntactically and semantically complementary to each other in such a way that the native speaker perceives them as a syntactic but also semantic continuum. In the case of simple collocations mapping onto verb argument structures, we cannot speak of a recursive feature but rather of *complementarity*, since complementary argument-specific collocations simultaneously combine with each other.

This is, for instance, the case of some usual constructions of the Italian verb *accompagnare* (EN *to accompany*) such as *il padre accompagna la sposa all'altare* (EN *the father walks the bride down the aisle*), *il cantante è accompagnato al pianoforte* (EN *the singer is accompanied on the piano*), or *piatti tradizionali accompagnati da ottimi vini* (EN *traditional cuisine accompanied by excellent wines*). It is clear that the verb and its arguments build a coherent *scene* (intended as a conceptual entity in the sense of Fillmore 1975) in which each component fulfils a cognitively functional role. Unlike simple collocations, these scenes can better reflect cultural specificities, with complex collocations, e.g. Italian *la madre accompagna i bambini a scuola* (EN *the mother takes her children to school*) possibly matching free, unremarkable word combinations in other languages.

Here, the main source of extension restriction is the valency of a verb, i.e. the number of elements within a complex collocation primarily depends on the number of arguments of the verb. The collocational range of simple collocations for each specific argument naturally becomes a further criterion for restriction.

Literature on collocations sometimes contains examples for complex collocations with argument complementarity without distinguishing it from the recursive expansion type. In describing German and Spanish specialised collocations in the field of investment funds, for instance, Ana Caro Cedillo (2004: 78) points out that “Die Zwei-Konzepte-Kollokation ist die Grundform. Einfache Kollokationen können aber weiter von anderen Elementen bestimmt werden. Sie können sich miteinander verketten und komplexe, aus mehr als zwei Konzepte bestehende Kollokationen bilden”[3]. Besides examples of recursively expanded collocations (e.g. *einen Wertzuwachs von x% erzielen*, EN *to achieve a x% increase in value,* ibid.: 215), the book also provides examples such as *dem Anteilwert einen Ausgabeaufschlag hinzuzurechnen* (EN *to add an issuance fee to the unit value*) as a complex collocation of *Ausgabeaufschlag* (EN *issuance fee*), which is a simple collocation in the form of a compound (ibid.: 223). All elements of the complex collocations are paired with arguments of the verb *hinzurechnen* (EN *to add*), which has a three-argument structure covering the syntactic functions subject, direct object, and indirect object).

2.3 Comparing modalities of complex collocation formation

The two modalities of complex collocation formation can be compared along four salient features which have been summed up in Table 1.

Formation:	Recursive expansion	Argument complementarity
a) Formation process:	Expansion of a binary collocation through the progressive addition of words	Concatenation of simple collocations of a verb matching two or more of its arguments
b) Semantic core:	Content word	Verb taking at least two arguments
c) Restriction rationale:	Collocational range	Valency
d) Conceptual range:	Phrase level	Sentence level

Table 1: Comparison between different modalities of complex collocation formation.

a) The *formation process* by which a complex collocation is created differs depending on the semantic core: a simple collocation can be gradually expanded by the addition of new lexical items or, alternatively, it can concatenate with other collocations corresponding to further arguments of the core, typically a verb.

b) The *semantic core* of a complex collocation built by recursive expansion can be any type of content word, for instance a noun modified by an adjective, a verb by an adverb, or an adjective by an adverb. In the case of complex collocations built by argument complementarity, the semantic core is typically constituted by a verb taking at least two arguments, e.g. a subject and a direct object[4]. The status of a word as a base or a collocate (cf. Hausmann 1985: 119) is not taken into account, since any element of a simple collocation can serve as the semantic core of the new collocation.

c) As previously mentioned in this section, the two formation modalities are also characterised by crucial differences in the *restriction rationale*. The primary principle behind the possibility that a simple collocation builds a complex one by recursive expansion is the collocational range of its elements, whereas in the case of argument complementarity a restraint is imposed by the number of arguments of the semantic core.

d) The *conceptual range* of a complex collocation is the syntactic level at which its concept is encoded. From this perspective, complex collocations built by recursive expansion have the same characteristics as the simple collocations from which they originate. A noun phrase, for instance, is expanded into a larger noun phrase by the addition of an adjectival modifier, or a verb phrase is expanded into a larger verb phrase by the addition of an adverbial modifier: in both cases, the concept encoded by the complex collocation is still specified at the phrase level. Concepts covered by argument-related complex collocations, on the contrary, are embedded at sentence level. This level is also able to identify complex *scenes* as shown in Section 2.2.

3 Complex collocations in pattern-based learner’s lexicography

Argument complementarity in complex collocations is crucial for meaning description in learner’s lexicography, providing learners with a consistent, non-fragmented view of the phraseological templates typical of a language. This is particularly true of dictionaries focusing on sense patterns based on argument structures, as described in Section 1. The Italian verb *inseguire* (EN *to chase*, *to pursue*), for instance, counts among its sense patterns the pattern *inseguire qualcuno (che cerca di non farsi raggiungere)* (EN *to chase someone (who is trying not to be caught up)*). Both obligatory

[3] “The two-concept collocation is the basic form. However, simple collocations can be further determined by new elements. They can be linked together to form complex collocations consisting of more than two concepts” (our translation).

[4] Content words other than verbs can also constitute semantic cores, as long as they have their own arguments. This is, for example, the case of the Italian nouns *libertà + di parola* (EN *freedom + of speech*) or *rispetto + delle regole* (EN *adherence + to the rules*).

arguments, subject phrase and object phrase, subsume a number of simple collocations. From the perspective of the sense pattern as a cognitively founded unit of meaning, however, some simple collocation pairs appear to be syntactically and semantically linked to each other and build complex collocations, e.g. *il cacciatore insegue la selvaggina* (EN *the hunter chases the game*).

We are now going to discuss some issues related to the extraction of this type of complex collocation from corpora, and their use in the microstructure of a pattern-based learner's dictionary.

3.1 Extracting complex collocations from a corpus

The analysis of significance scores, and possibly of parsed data, is useful for validating collocation candidates and could potentially be applied to any kind of collocation. However, widespread corpus query systems primarily concentrate on the extraction of simple collocation candidates, while the retrieval of complex collocations is usually left to the linguistic and technical skills of the lexicographer.

For the identification of complex collocations we employ different methods and tools. We first use Sketch Engine (Kilgarriff et al. 2004), collecting collocations of a node verb from the Italian Web 2016 corpus through the Word Sketch tool. This tool extracts binary word combinations, sorting them according to the specific grammatical relations defined by a *sketch grammar*. The corpus needs to be POS-tagged and lemmatised, whereas no parsing is required. Whenever relevant, Word Sketch displays the most frequent representation of a binary combination, possibly revealing some complex collocation. Table 2 shows some results for the Italian verb *inseguire* (EN *to chase, to pursue*), with the example of a complex collocation candidate based on argument complementarity:

Search word:	Collocation candidate:	Most frequent form in the corpus:
inseguire (EN *to chase, to pursue*)	subject of *inseguire*: *squadra* (EN *team*)	*due squadre si inseguono* (EN *two teams chase each other*) (binary candidate)
	subject of *inseguire*: *notte* (EN *night*)	*la notte insegue sempre il giorno* (EN *the night always follows the day*) (complex candidate with argument structure: subject + direct object)

Table 2: Search for complex collocations in the Italian Web 2016 corpus through word sketches (Sketch Engine).

This is by no means an efficient solution for identifying complex collocations in the corpus. In order to systematically look for argument-related complex collocations, we use the Sketch Engine's multiword sketch function, which extends the search for collocation candidates to further collocates of the original word sketches. This expansion enables the detection of complex collocations, as exemplified by the Italian noun *braccio* (EN *arm*) and the verb *accompagnare* (EN *accompany*) in Table 3:

Search word:	Collocation candidate:	Complex collocation candidate:
braccio (EN *arm*)	verbs with *braccio* as object: *tendere* (EN *stretch*)	modifiers of *tendere* + *braccio*: *destro* (EN *right*) (recursively built complex candidate: *tendere il braccio destro*, EN *to reach out your right arm*)
accompagnare (EN *to accompany*)	object of *accompagnare*: *visitatore* (EN *visitor*)	subject of *accompagnare* + *visitatore*: *guida* (EN *guide*) (argument-related complex candidate: *la guida accompagna il visitatore*, EN *the guide accompanies the visitor*)
		prepositional phrase with *accompagnare* + *visitatore*: *lungo il percorso* (EN *along the route*) (argument-related complex candidate: *accompagnare il visitatore lungo il percorso*, EN *to accompany the visitor along the route*)

Table 3: Search for complex collocations in the Italian Web 2016 corpus through multiword sketches (Sketch Engine).

Candidates are then validated by considering frequency and score of each collocation, and by introspection, in particular through the analysis of collocation contexts within GDEX-filtered corpus samples (Kilgarriff et al. 2008) and the comparison with data from general and collocation dictionaries. The procedure that needs to be followed in order to find and validate complex collocates requires a considerable amount of time, which is mainly due to the fact that searches have to be separately performed on each simple collocation candidate. Experiments carried out with German corpora provided by the DWDS Wortprofil tool (http://dwds.de/d/wortprofil) substantiate these observations. Moreover, there seems to be no correlation between the availability of complex collocations at the level of sense patterns and the semantic specificity of a verb: in fact, we did not notice any particular difference between semantically generic verbs and more specific verbs.

Another method for retrieving complex collocations is the use of corpus query languages to formulate complex queries, including multiple arguments of a verb. We carried out concordance searches employing the Corpus Query Language option in Sketch Engine, and also performed some tests on different corpora using the Corpus Query Processor provided by the IMS Corpus Workbench (Evert & Hardie 2011). The main limits of this method lie in the mandatory predefinition of the specific argument structure of a verb, as well as in the lack of a specific evaluation frame for collocation significance.

To the best of our knowledge, the topic of extraction of complex collocations from corpora has been treated only marginally in relevant literature. In the case of the extraction of adjective-noun-verb combinations, Zinsmeister & Heid (2003) pleaded for a parsing procedure with a lexicalised probabilistic grammar instead of a simple pattern-matching on part-of-speech shapes. A parsing-oriented, syntax-based approach to collocational data extraction is also discussed by Seretan (2011), who proposes the pre-processing of extracted bigrams in order to automatically infer longer collocations (ibid.: 103 ff.). This method identifies recursively built nested collocations such as *treaty on the non-proliferation of weapons of mass destruction* (ibid.: 104). Despite general usefulness of this type of complex collocation, the results are not sufficient for the purpose of sense pattern description.

The method proposed by Kraif & Diwersy (2014) also relies on a parsed corpus, from which *lexicograms*, i.e. models for the main syntactic collocates of a given node together with association measures, are extracted and can be recursively employed to find increasingly longer combinations. The definition of specific syntagmatic structures allows, in this case, for more precise results in terms of argument structures. For the French node noun *respect* (EN *respect*) in a verb-object relation, for instance, the output would include *inspirer un profond respect* (EN *to command deep respect*) and *imposer le respect des normes* (EN *to enforce compliance*), which would match the argument structure of the input noun. However, it is not clear to what extent extraction from a corpus can be carried out in a systematic way for the complete range of arguments of any search word.

From a lexicographic standpoint, the automated extraction of argument-related complex collocations from corpora still presents general problems in terms of time efficiency, coverage, and quality of results. Section 4 will mention some interesting perspectives for future research on this topic.

3.2 Complex collocations in a pattern-based learner's dictionary

After candidate validation, complex collocations are employed for different tasks in the dictionary making process. Not only do they play a significant role in the metalexicographic activity of sense pattern formulation, as illustrated in Section 3.2.1, they are also intended to be recorded as lexicographic data in dictionary entries. At the level of the dictionary's microstructure, in fact, the presentation of complex collocations can primarily take place in the following mutually exclusive ways:

- complex collocations as semantic types or semantic roles (Section 3.2.2);
- complex collocations as subordinate treatment units (Section 3.2.3).

3.2.1 Complex collocations as a base for sense pattern formulation

Complex collocations extracted from a corpus and manually validated are systematically employed for the formulation of sense patterns. This essential metalexicographic function is fulfilled in two ways. On the one hand, analysing and grouping syntactically and semantically homogeneous collocations supports the identification of a specific sense pattern, i.e. of a distinct argument structure associated with a distinct meaning. For instance, the complex collocations *il poliziotto insegue il ladro* (EN *the policeman pursues the thief*) and *il malintenzionato insegue la vittima* (EN *the ill-intentioned person pursues the victim)* help identify the sense pattern *inseguire qualcuno (che cerca di non farsi raggiungere)* (EN *to chase someone (who is trying not to be caught up)*)[5].

On the other hand, the analysis of paradigmatic structures matching verb arguments supports the selection of appropriate semantic types and semantic roles needed for the formulation of sense pattern. Semantic types, together with semantic roles, are the fundamental meaning components in sense patterns: in a hierarchy of concepts, they lexically represent the least common subsumer of all lexical items matching a specific verb argument in a specific verb meaning (Hanks 2004, DiMuccio-Failla & Giacomini, 2017a). The generic noun *vehicle*, for instance, is the suitable semantic type for a large cluster of lexical items such as *car*, *truck*, *bicycle*, *train*, or *ship*.

Observing collocate paradigms such as the ones at the subject and direct object levels in *il poliziotto/l'agente/la pattuglia/... insegue il ladro/il criminale/il sospettato/...* (EN *the policeman/the cop/the patrol/... pursues the thief/the criminal/the suspect/...*) or *il malintenzionato/il criminale/... insegue la vittima* (EN *the ill-intentioned person/the criminal pursues the victim)* is central for this process. The direct object of *inseguire* (EN *to pursue*) in this particular meaning, for instance, has been associated with the semantic type *qualcuno (che cerca di non farsi raggiungere)* (EN *someone (who is trying not to be caught up)*), which displays the most suitable level of generalization for subsuming all available collocates.

Of course, adequate semantic types do not always match the collocate of a verb, and sometimes they can even coincide with quite uncommon concepts, as exemplified by the English verb *toast*, with *breadstuff* as a semantic type for usual direct objects such as *bread*, *bun*, or *sandwich* (see discussion in DiMuccio-Failla & Giacomini: 2017a).

The following examples illustrate sense patterns with some of their subsumed complex collocations:

[5] Complex collocations extracted from corpora may involve adjuncts, e.g. *a piedi* (EN *on foot*) in *il poliziotto insegue il ladro a piedi* (EN *the policeman pursues the thief on foot*). However, it needs to be pointed out that in our lexicographic model, sense patterns of verbs typically cover syntactic and semantic arguments, but no adjuncts. Adjuncts are therefore not used for pattern formulation.

i) IT seguire (la direzione indicata da) una cosa (EN *to follow (the indication given by) something*)
--subsumes--> *seguire la propria vocazione* (EN *to follow one's vocation*) (complex collocation)
--subsumes--> *seguire il richiamo della foresta* (EN *to follow the call of the wild*) (complex collocation)

ii) IT pedinare una persona (che sta andando da qualche parte) (EN *to tail someone (who is going somewhere)*)
--subsumes--> *il poliziotto pedina il sospettato* (EN *the policeman is tailing the suspect*) (complex collocation)
--subsumes--> *il malintenzionato pedina la vittima* (EN *the attacker is stalking the victim*) (complex collocation)

iii) IT guidare un veicolo in un certo luogo (EN *to drive a vehicle to a particular place*)
--subsumes--> *guidare la macchina fino al parcheggio* (EN *to drive the car to the parking lot*) (simple collocation only: *guidare la macchina*)
--subsumes--> *guidare una nave in porto* (EN *to steer a ship into harbor*) (complex collocation)

iv) DE *einen Befehl (insb. einer Autorität) befolgen* (EN *to obey an order (esp. from an authority)*)
--subsumes--> *den Befehl des Vorgesetzen befolgen* (EN *to obey one's superior's orders*) (complex collocation)
--subsumes--> *der Soldat befolgt das Kommando* (EN *the soldier obeys the command*) (complex collocation)

3.2.2 Complex collocations as semantic types or semantic roles

In some cases verb collocates already reaching the most suitable level of generalisation can be directly elevated to semantic types, as shown in the following sense pattern examples for Italian and German[6]:

v) IT guidare un veicolo (EN *to drive a vehicle*) (simple collocation: verb – direct object)
vi) IT tallonare un avversario in una classifica (EN *to tail an opponent in a ranking*) (complex collocation: verb – direct object – adverbial)
vii) DE eine Aktivität wird von (dem Klang einer/eines) Stimme/Musikinstrument(s) begleitet (EN *an event is accompanied by (the sound of) a voice/music instrument*) (complex collocation: subject – verb – direct object)

The main advantage of presenting semantic types or roles of sense patterns by means of complex collocations lies in the fact that dictionary users are simultaneously provided with typical scene-like structures and typical phraseological units. As is always the case for semantic types, complex collocations with this function may encompass more specific lexical items, among which are further collocates of the verb:

viii) DE eine Aktivität wird von (dem Klang einer/eines) Stimme/Musikinstrument(s) begleitet (EN *an event is accompanied by (the sound of) a voice/music instrument*)
--subsumes--> *der Sonnenaufgang wird vom Gesang der Vögel begleitet* (EN *the sunrise is accompanied by the song of the birds*) (simple collocation only: *Gesang der Vögel*)
--subsumes--> *der Sänger wird am Klavier begleitet* (EN *the singer is accompanied on the piano*) (complex collocation)

3.2.3 Complex collocations as subordinate treatment units

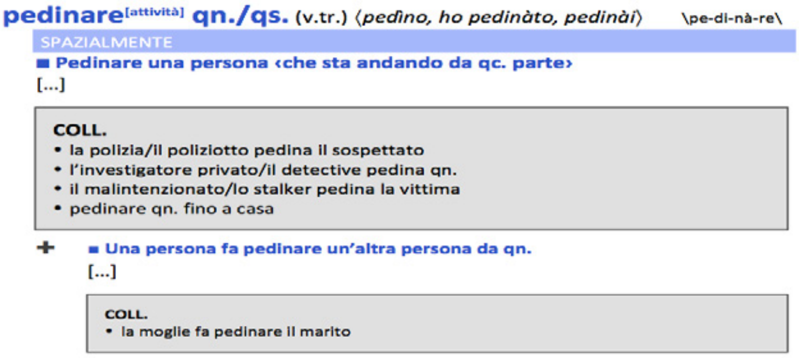
Complex collocations that are not selected for presentation within a sense pattern, e.g. the subsumed collocations mentioned in examples i-iv and viii, can still be allocated in a subordinate treatment unit (cf. Gouws 2015) with a distinct search zone for each sense of the verb[7]. In the entry of the verb *pedinare* (EN *to tail*), the sense pattern *pedinare una persona (che sta andando da qualche parte)* (EN *to tail someone (who is going somewhere)*) as a treatment unit can be presented as in Figure 1[8].

Both the given sense pattern and its subpattern *una persona fa pedinare un'altra persona da qualcuno* (EN *someone has someone else being followed by a person*) build independent microstructural treatment units that include a subordinate component dedicated to collocations. Being compositional phraseological units, complex and simple collocations do not require a meaning paraphrase.

[6] The sense pattern is underlined in each example.

[7] We deliberately use the adjective *subordinate* instead of *secondary* to distinguish it from the traditional vision of the lemma as the *primary* treatment unit (Wiegand 1996, Gouws 2015), highlighting at the same time the dependence of collocations on sense patterns as the superordinate and key microstructural element of our pattern-based dictionary model.

[8] Features illustrated in Figure 1 and Figure 2 are excerpts from a prototype of the planned pattern-based dictionary.



pedinare[attività] **qn./qs.** (v.tr.) 〈*pedìno, ho pedinàto, pedinài*〉 \pe-di-nà-re\

SPAZIALMENTE

■ **Pedinare una persona ‹che sta andando da qc. parte›**

[...]

COLL.
- la polizia/il poliziotto pedina il sospettato
- l'investigatore privato/il detective pedina qn.
- il malintenzionato/lo stalker pedina la vittima
- pedinare qn. fino a casa

✚ ■ **Una persona fa pedinare un'altra persona da qn.**

[...]

COLL.
- la moglie fa pedinare il marito

Figure 1: Excerpt from the entry for the Italian verb *pedinare* (EN *to tail*) with a sense pattern and a sense subpattern.

In discussing the insertion of complex collocations in dictionaries, Gouws (2015: 184-185) states the following:

> The inclusion of complex collocations remains important and lexicographers should negotiate the best possible way of presenting them and of making users aware of their existence. This could be done either as guiding elements in a subcategory of the search zone for collocations or in a more implicit way as part of the treatment of single collocations, typically within an example sentence illustrating the use of the single collocation but also its occurrence as component of a complex collocation.

Whereas recursively built complex collocations can be easily seen as a subcategory of simple collocations, we think that argument-related complex collocations should be treated as a superordinate category subsuming simple collocations.

The planned data representation in XML format allows for a hierarchical distribution of microstructural items (Figure 1) and for the attribution of argument-related collocates in the collocation treatment unit to the arguments of the corresponding sense pattern (Figure 2). Dictionary users will then be able to search for complex collocations matching all arguments or simple collocations matching specific arguments of a verb pattern. Adding thematic roles to argument representation further increases the degree of semantic detail. Whenever relevant, results can be expanded to adjunct-related collocations, e.g. *pedinare qualcuno fino a casa* (EN *to tail someone all the way to their home*). Figure 2 shows all validated collocates of *pedinare* identified for the selected sense pattern: collocates are sorted according to the related argument and, in the case of complex collocations, are linked to further collocates, e.g. in *il poliziotto pedina il sospettato* (EN *the policeman is tailing the suspect*).

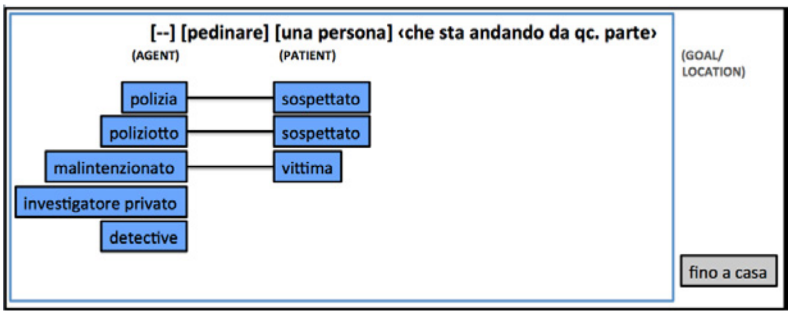


Figure 2 – Prototype visualization of results for search queries combining argument structures, thematic roles and collocates of the Italian verb *pedinare* (EN *to tail*): simple and complex argument-related collocations are highlighted in blue, adjuncts in grey.

This data model highlights complex collocations as the phraseological building blocks of sense patterns, and simple collocations as their basic constituents. Moreover, any search parameter can be employed here by the user to perform a query and be combined with other parameters, e.g. to retrieve all complex collocations for a given sense pattern, all complex collocations matching specific arguments, or all collocates matching a specific thematic role.

Embedding argument-related complex collocations in lexicographic examples provided to illustrate the use of a sense pattern is an option we generally do not take into account, since the information conveyed by this solution might be too implicit for language learners.

4 Conclusions

This contribution has focused on a type of complex collocation involving the level of argument structures, in particular of verbs, showing how they can support multiple metalexicographic and lexicographic functions in a pattern-based learner's dictionary. Initial tests for the evaluation of complex collocation coverage within the planned dictionary were recently carried out at Heidelberg University by MA students in translation with Italian or German as their native (L1) or second/foreign language (L2/FL). Pattern-specific collocations of selected polysemous verbs were employed for performing a text reception and an active translation task. Participants were asked (1) to identify sense patterns of the same verb in a L2/FL text by analysing the available collocations, and (2) to translate L1 sentences containing different senses of the same verb into the L2/FL by finding the right sense pattern through collocations. First results show that simple collocations of verbs are usually exhaustive enough to support sense disambiguation during text reception, whereas sense disambiguation for text production in (and translation into) the foreign language is improved by the availability of complex, argument-related collocations. These results seem to confirm the status of argument-related complex collocations as phraseological lexicalisations of conceptual scenes, and thus their relevance for language learners. Further tests on new datasets are planned for the future.

Some relevant issues have been highlighted in the paper regarding the extraction of complex collocations from corpora. Our future research on this topic will explore current findings in the field of terminology extraction, which can possibly yield some insights into methods for the detection and validation of n-grams and term variants in corpora. An interesting perspective in this context is also the investigation of the possibilities opened up by neural word embeddings, in particular in the field of analogy recovery (cf. Goldberg 2017). As pointed out in Section 3, reliable results are needed in terms of time efficiency, coverage, and quality of extracted data. However, we are convinced that only a solid underlying theory on complex collocations can support the development of new lexicographic-oriented procedures.

5 References

Almela-Sánchez, M. (2019). Collocation and Selectional Preferences: A Frame-based Approach. *Journal of English Studies*, 17 (2019), pp. 3-41.

Alonso Ramos, M. & Wanner, L. (2007). Collocation chains: how to deal with them? In K. Gerdes, T. Reuther, L. Wanner (Eds.), *Proceedings of the Third International Conference on Meaning-Text Theory* (pp. 11-20). Wiener Slawistischer Almanach, Sonderband 69. Munich.

Cedillo, A. C. (2004). *Fachsprachliche Kollokationen: Ein übersetzungsorientiertes Datenbankmodell Deutsch-Spanisch* (Vol. 63). Tübingen, Germany: Gunter Narr Verlag.

DiMuccio-Failla, P.V. & Giacomini, L. (2017a). Designing an Italian learner's dictionary based on Sinclair's lexical units and Hanks's corpus pattern analysis. In *Proceedings of the Fifth eLex Conference Electronic Lexicography in the 21st Century*. Leiden, Netherlands.

DiMuccio-Failla, P.V. & Giacomini, L. (2017b). In M. Mitkov (ed.), *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017*, LNAI 10596. Springer, pp. 290-305.

Evert, S. & A. Hardie (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. http://cwb.sourceforge.net/index.php

Fillmore, C. J. (1975). An alternative to checklist theories of meaning. In: *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, pp. 123-131.

Geeraerts, D. (2007). Lexicography. In: D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics*. OUP, pp. 1160-1174.

Giacomini, L. & DiMuccio-Failla, P. (2019). Investigating Semi-Automatic Procedures in Pattern-Based Lexicography. In: *Proceedings of the eLex 2019 conference. Electronic lexicography in the 21st century.* Sintra, Portugal.

Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

Gouws, R. H. (2015). The presentation and treatment of collocations as secondary guiding elements in dictionaries. *Lexikos*, *25*, pp. 170-190.

Gross, G. (1994). Classes d'objets et description des verbes. *Langages*, pp. 15-30.

Halliday, M. A. K. (1992). Some lexicogrammatical features of the zero population growth text. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Benjamins, pp. 327-358.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.

Hanks, P. (2004). Corpus Pattern Analysis. In: *Proceedings of the XI EURALEX International Congress*, Vol. 1, 87-98.

Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In: H. Bergenholtz & J. Mugdan (Eds.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch* 28. - 30. 6. 1984. Niemeyer, pp. 118-29.

Heid, U. (1994). On ways words work together – research topics in lexical combinatorics. In: *Proceedings of the VI EURALEX International Congress*, Amsterdam, pp. 226–257.

Herbst, T. (2017). Menschliche Sprache: Ein Netzwerk aus Mustern genannt Konstruktionen. *Sprachwelten: Vier Vorträge. Erlanger Universitätstage*, pp. 105-147.

Herbst, T., Schmid, H. & Faulhaber, S. (Eds.) (2014). *Constructions Collocations Patterns*. De Gruyter Mouton.

Herbst, T. (2016). Wörterbuch war gestern. Programm für ein unifiziertes Konstruktikon! In: S. Schierholz, R. H. Gouws, Z. Hollós & W. Wolski (Eds.), *Wörterbuchforschung und Lexikographie.* De Gruyter, pp. 169-206.

Herbst, T. (2018). Is language a Collostructicon? – A Proposal for Looking at Collocations, Valency, Argument Structure and Other Constructions. In: P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis: Advances and Applications.* Springer, pp. 1-22.

Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English* (Vol. 4). John Benjamins Publishing.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In: *Proceedings of the XIII EURALEX International Congress*, Barcelona, pp. 425-431.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 The sketch engine. *Information Technology*, pp. 105-116.

Kraif, O. & Diwersy, S. (2014). Exploring combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions. In: P. Blumenthal, I. Novakova & D. Siepmann (Eds.), *Les émotions dans le discours.*

Emotions in discourse. Peter Lang, pp. 381-394.

McIntosh, A. (1966). Patterns and ranges. *Language* (37), pp. 325-337.

Ostermann, C. (2015). *Cognitive lexicography: A new approach to lexicography making use of cognitive semantics*. de Gruyter.

Renau, I. & Nazar, R. (2016). Automatic Extraction of Lexical Patterns from Corpora. In T. Margalitazde & G. Meladze (eds.) In *Proceedings of the XVII EURALEX International congress. Lexicography and linguistic diversity*, pp. 823-830.

Seretan, V. (2013). A multilingual integrated framework for processing lexical collocations. *Computational Linguistics.* Springer, pp. 87-108.

Seretan, V. (2011). *Syntax-based collocation extraction*. Springer Science & Business Media.

Sinclair, J. McH. (1996). The search for units of meaning. *Textus: English Studies in Italy* 9(1), pp. 75-106.

Sinclair, J. McH. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2), pp. 209-243.

Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language 2,* pp. 23-55.

Tutin, A. & Kraif, O. (2016). From binary collocations to grammatically extended collocations: Some insights in the semantic field of emotions in French. *Mémoires de la Société néophilologique de Helsinki, Helsinki: Société néophilologique de Helsinki, 2016, Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching, hal-01337486*, pp. 245-266.

Wehrli, E., Seretan, V. & Nerima, L. (2010). Sentence Analysis and Collocation Identification. In: *COLING 2010, Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, Beijing, pp. 28-36.

Wiegand, H. E. (1996). Das Konzept der semiintegrierten Mikrostrukturen. Ein Beitrag zur Theorie zweisprachiger Printwörterbücher. In: H E. Wiegand (Ed.), *Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographischen Kolloquium*. Tübingen (Lexicographica. Series Maior 70), pp. 1–82.

Zinsmeister, H., & Heid, U. (2003). Significant triples: Adjective+ noun+ verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003)*, Budapest.