Constructions, Collocations, and Patterns: Alternative Ways of Construction Identification in a Usage-based, Corpus-driven Theoretical Framework

Gábor Simon
Eötvös Loránd University Budapest
simon.gabor@btk.elte.hu

Abstract

There is a serious theoretical and methodological dilemma in usage-based construction grammar: how to identify constructions based on corpus pattern analysis. The present paper provides an overview of this dilemma, focusing on argument structure constructions (ASCs) in general. It seeks to answer the question of how a data-driven construction grammatical description can be built on the collocation data extracted from corpora. The study is of meta-scientific interest: it compares theoretical proposals in construction grammar regarding how they handle co-occurrences emerging from a corpus. Discussing alternative bottom-up approaches to the notion of construction, the paper concludes that there is no one-to-one correspondence between corpus patterns and constructions. Therefore, a careful analysis of the former can empirically ground both the identification and the description of constructions.

1 Introduction

If there is a dichotomy between construction grammar and NLP technologies, it can be considered a multidirectional theoretical problem as well. On the one hand, NLP models and methods need constant theoretical support from CxG approaches to language. On the other hand, constructionist frameworks also need to devote more attention to data extraction techniques, because this may result in a more appropriate bottom-up modelling of the complex system of the constructicon. The present paper aims at bridging the gap between data-driven collocation analysis and the theoretical endeavor of construction grammar, focusing on argument structure constructions (ASCs). Thus, the following pages (merging the genres of a metatheoretical proposal and a critical review) provide the reader not with an empirical analysis of a specific issue, but with an overview of how we can refine our knowledge of constructions based on collocation patterns.

In the most general sense, grammatical constructions are form-meaning pairs that are at least partially arbitrary (Croft 2001: 18), i.e., some aspect of their form or function cannot be predicted from their components or from other existing constructions (Goldberg 2006: 5). Building on this definition, constructionist grammatical approaches model our knowledge of the language as a network of constructions of varying complexity: morphemes, word forms and syntactic structures (Goldberg 2019: 36; Croft 2001).

This kind of fluidity and flexibility certainly has a liberating effect on theorizing, since a single concept can explain an extremely wide range of phenomena previously isolated into rigid taxonomies. However, it may paralyse corpus-driven research based on data analysis, since it does not even provide the researcher with clear concepts for defining and/or identifying the central phenomenon. What should be the size (and complexity) of the structure whose occurrences are to be analyzed? How large a sample should we take? What quantifiable data will be relevant in mapping the diversity of the phenomenon?

Illustrating the emerging problems with a specific example, consider the following issues: is the noun of the expression *kick the bucket* a construction in its own right, or can it only be

Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023), pages 12 - 20
March 9-12, 2023 ©2023 Association for Computational Linguistics

described as a component of the construction as a whole? If we accept the former, what is the relation between *bucket* and the nouns in the expressions *kick the ball* (in a soccer match) or *kick the habit*?

Moving a step further, is the structure *kicked the bucket* merely a realization of the initial example above, or is it an independent construction, given that in the COCA corpus the past tense verb form has almost the same frequency (55) as the infinitive (70), much more than the present tense singular third person form of the verb (19)?[1] And to what extent can a CxG analysis distinguish between the structure *kicking the bucket* and the structure *emptying the bucket*, if the collocation strength of the two verb forms does not differ significantly (7.85 and 8.83 in MI score)?

William Croft (2001: 17) summarizes this dilemma in an illuminating way: "[t]he constructional tail has come to wag the syntactic dog: everything from words to the most general syntactic and semantic rules can be represented as constructions." This leads us to the following questions: (i) How can the researcher delineate individual constructions as empirical facts in language use? (ii) How can data-driven analysis of corpus patterns support construction grammatical description?[2]

A possible solution to the problem of construction identification in a corpus may lie in the adaptation of the distributional approach to construction grammar (Goldberg 2019: 39): to decide what will be a construction in a language, we must first identify the units that express the same thing in a similar or identical way, then observe their distribution and what other constructions might belong to that category. However, the main assumption of a distributional analysis is that there is invariability either in meaning or in form. Since constructions are holistic representations, considerable formal differences (e.g., person or tense marking) may instantiate the same schematic construction, while relatively small modifications in the form (e.g., replacing a nominal argument with another) may lead to a new construction.

One problem obviously arises from the predetermination of either the form or the meaning without having observed the data themselves. Our decision can only be theoretical, which then either works on a wide range of data (with the risk of overgeneralization and the loss of explanatory power) or necessarily narrows the scope of the construction analysis. Another one comes from the analysis of overlapping distribution. Is it the morphological elaboration of verb forms? Does it extend to word order? Or to the presence or absence of additional (potential) arguments? In other words: how many details do we need to take into consideration when describing the variability of a hypothetical construction to draw conclusions from the data at a higher level of abstraction?

Again, these questions cannot be answered from the perspective of construction grammar, which presupposes a usage-based approach in which different levels of generalizations constitute our knowledge about language. Goldberg (2006: 64), for example, defines the essence of a usage-based approach as taking into account both the facts of the actual use of linguistic expressions (frequencies, specific patterns) and the cases of generalizations (schema-level knowledge). That is, in addition to instance-based representations, knowledge of more generic constructions is also assumed, and the network of the constructicon is therefore multilevel. (See also Croft 2001: 25 and Bybee 2013 on further claims of a usage-based construction grammatical approach.) Consequently, there is no distinguishing feature which, if observed, makes the distinction between constructions clear.

It is instructive how Croft (2001: 28) formulates this dilemma: "the degree of generality of construction schemas, and the location of grammatical information in the taxonomic network is an empirical question to be answered by empirical studies of frequency patterns and psycholinguistic research on entrenchment and productivity of schematic constructions". Nevertheless, to extract assumed constructions from a data set, we need to posit the construction beforehand.

The data type of collocations may seem a good candidate for a data-driven construction identification. However, we do not know to what extent collocations can be considered constructions in themselves, or to what extent they can be used as

[1] The data are from the COCA corpus (https://www.english-corpora.org/coca/), last access: 11/09/2022.

[2] For the sake of clarity, it is worth noting that the present study focuses only on corpus-driven methodological framework. Therefore, corpus-based and/or corpus-assisted investigations are not the target of the paper.

a parameter for describing a construction. Consequently, collocations cannot be considered a priori data for construction identification, therefore any corpus analysis needs to determine in advance what kind of construction it wants to explore. This in turn may make scientific reasoning more circular. The primary aim of the present paper is to provide the reader with alternative ways out of this circularity.

To summarize the theoretical and methodological dilemmas raised here, we can conclude that the conceptualization of the construction and the multi-level network model of construction grammars are not very conducive to systematic and data-driven corpus analyses. As Thomas Herbst and his colleagues note, "while many usage-based researchers in cognitive linguists have, of course, embraced the corpus method, it is still true to say that they have been more interested in arriving at generalizations than in reaching the level of descriptive granularity and specificity that is typical of more traditional corpus-based approaches" (Herbst et al. 2014: 4).

In the following, first I outline the possibilities and limitations of using data types of corpus linguistics in construction descriptions (2). Then those proposals building on collocation-like patterns for mapping constructions are discussed (3). The paper ends with concluding remarks (4).

2 How can corpus data help to identify constructions?

This section aims to provide a brief outline of those corpus linguistic data types that may ground the analysis of construction based only on observable facts of language use. As Stefan Gries (2013) points out, the inclusion of corpus linguistic tools in the description of constructions is a significant shift from the early introspective methods of construction linguistic analysis. The simplest data is if there is no data, i.e., the lack of any occurrence of a construction in the corpus. It serves as an argument against the hypothetical existence of the construction based on intuition. Starting from the absence of occurrence as an extreme case, the corpus provides two types of data for construction identification: frequency and co-occurrence. However, the question is not only how and by what means we measure and make these phenomena observable, but also how we interpret them.

Absolute frequency, the total number of occurrences of a unit in the corpus, proves to be informative when one wants to observe the central variants of a structure. For example, the most frequent verb + argument combinations for each verb, or the arguments most frequently realized with a verb. The methodological limitations of this data type stem from the fact that the calculation of the absolute frequency assumes a prior definition of the unit to be measured.

Compared to the number of occurrences, the co-occurrence rate, i.e., the degree to which two (or more) words are associated in the corpus seems to be more informative. The most familiar category of co-occurring words is the one of collocation. Two words are collocated if their association is statistically significant. Collocation extraction can be performed with two kinds of method (Seretan 2011: 3): according to the n-gram method, a sequence of consecutive words can be considered fixed units of collocability (therefore, n-grams shed light on fixed word-order patterns); whereas the window method takes the context in a broader sense and explores all potential collocates that typically occur in the corpus within a certain context of the node.

Beyond counting the number of co-occurrences of words in a corpus, the other aspect of collocability is the strength of co-occurrence patterns. It can be measured using different association scores (see Evert 2009; Levshina 2015: 234–235 for more details). Without going into details about the different methods of calculation, it is worth pointing out that each value highlights a different aspect of the observed patterns. For example, the Mutual Information (MI) score is sensitive to fixed lexical units (e.g., names, phraseological units, idioms) and favors infrequent terms, the so-called hapaxes. Therefore, MI measures are particularly useful for lexicography. Other values (e.g., log likelihood, χ2 score or t-score) tend to make frequent grammatical patterns observable (Evert 2009: 1230). Thus, both idiomatic and more schematic constructions might be identified with the help of collocation extraction.

The association scores can also be distinguished according to their directionality: for instance, the ΔP value is unidirectional (i.e., the node associates the collocate or the collocate associates the node), while most of the values are bidirectional (i.e., they demonstrate a mutual association between members of the collocation). Even though ΔP is unidirectional, it is suitable for constructional measures (see Gries 2013), because one version of it can be used to measure association from the verb (ΔP, verb as cue, construction as response), and another version of it can give us data about the attraction of verb lexemes from the perspective of the construction (ΔP, construction as a cue, collexeme as response, Levshina 2015: 234). Therefore, directionality plays an important role in distinguishing the specific and schematic parts of a co-occurring pattern.

From this overview, it is perhaps clear that the observation of collocability may lead to a rich variety of verb + argument associations. But the question of whether these are real constructions remains open, which is why a more detailed methodological grounding is needed for this type of analysis. Indeed, the fact of collocability tells us how typical the occurrence of other words is in the narrower or wider context of a verb, but the reason for the occurrence of such word combinations, i.e. whether there is indeed a constructional behavior in the background or not, cannot be explained from the collocation data themselves. Seretan (2011: 4) argues that even if a pair of words does indeed typically occur together within a particular window, it is not certain that they are truly syntactically related terms, rather than random juxtapositions or mere noise (e.g., occurrences separated by a clause boundary or additional terms). Barnbrook et al. (2013: 164) draw a similar conclusion: collocations, despite their apparent significance as data type, are not really integrated into linguistic modelling.

The main problem of collocation measurement for constructional grammar is, therefore, that collocations themselves are not transparent in terms of constructions. Thus, just as we do not arrive at the empirical identification of constructions from the theoretical definition of the concept, we do not arrive at the identification of constructions on the basis of data types provided by the corpus. By way of explanation, there is no one-to-one correspondence between a pattern in a corpus and the concept of construction. Corpus analysis can help us to describe verb constructions with a variety of data, explore the features of the verbal components in them (via frequency patterns), identify fixed or flexible word order patterns (n-grams), reduce our effort to measure the variability of the construction by statistical measurements, and increase the efficacy of observing the variability of a given construction (collocation analysis). But the question of what counts as a construction in the corpus data remains unanswered even in quantitative analysis. As a consequence, for a corpus-driven description of constructions, it is necessary to narrow the gap between CxGs and corpus linguistics. In the following section, I present alternative theoretical proposals for such an attempt at integration.

3 Collocation-based construction analysis: alternative proposals

Three alternative theories of construction grammar that attempt to link the notions of construction and collocation are discussed here: radical construction grammar, the valence-based construction approach, and pattern grammar. While these theories initiate collocation-based construction identification in different ways, a common point is that extracting collocation patterns serves as the initial step to exploring higher levels of argument structure constructions.

3.1 Collocational dependencies

Croft (2001: 176-185) presents an analysis in which he considers two types of dependencies: coded and collocational dependencies. As shown in examples (1a-b), these relations are essentially syntactic in nature.

(1a) I have folks like you to open my eyes to see that love is weird, love is strange, love is good
(1b) Every time I open my eyes she is looking down at me.[3]

In both cases, the verb *open* has a subject and an object argument. The reflexive usage of the verb in (1b) demonstrates, however, that the process of opening the eye instantiates differently. In the first case, the multi-word unit can be interpreted as ‘to see the truth’, but in the second case, the meaning of the structure is ‘open the eyes/begin to see’. The examples thus show that the encoded and collocational dependencies do not coincide (despite all apparent similarities).

Similar observations led Croft to define collocational dependencies, which prescribe

[3] The data are from the COCA corpus (https://www.english-corpora.org/coca/), last access: 11/09/2022.

specific phrases besides the verb (e.g. the structure *into flower* in the context of the verb *burst*) or a group of phrases (e.g. the lemmata of *cherry tree/almond tree/fruit tree* etc. in the context of the phrase *burst into flower*), as symbolizing semantic rather than syntactic relations. The figure below summarizes this interpretation, using the English idiom *spill the beans* as an example.

The collocational relationship thus links two concepts at the semantic pole (e.g. *open* → MAKES TO SEE, *eyes* → TRUTH, cf. Figure 1), and the association observed in the corpus stabilizes these semantic correspondences in language use.

On this basis, collocational relations are inherently semantic in nature, which are then represented to varying degrees in a syntactically transparent way. Consequently, Croft postulates a continuum from purely semantic collocational relations through syntactically encoded collocations to those collocations that are not transparent in any way. For instance, the verb *blossom* has the following stable collocations in the COCA corpus: *flower* (6.43), *tree* (4.48), *rose* (5.50), and *garden* (3.52). [4] These collocations imply a selectional restriction, according to which the verb under consideration is combined with words referring to flowering plants (individually or in a group). In other words, the selectional restrictions that can be identified through collocations are purely semantic collocational dependencies that help to identify constructions. However, among the collocations, one can observe *romance* (7.09), *relationship* (3.54), *friendship* (6.47) or *career* (4.18). Since the latter violate the selectional restrictions emerging from the previously observed group of collocations, it follows that we are dealing here with another construction with a figurative meaning, in which the verb means ‘increases in intensity, unfolds vigorously’. Selectional restrictions are therefore not directly encoded syntactically, but they do help to identify the constructions organized around the verbs, and as collocational dependencies, they allow analyses based on word combinations.

Compared to purely semantic collocational dependencies, collocations proper represent a shift towards syntactic transparency. In Croft’s system (Croft 2001: 180) collocations proper listed above function as lower-level constructions and can be subordinated to more general constructions. The occurrences of *blossom* + *flower/tree/rose/garden* (etc.) are instances of the construction [*blossom* PLANTNOUN], while the data of *blossom* + *love/friendship/career/relationship* are instances of the construction [*blossom* PROCESSNOUN]. This is a productive approach because we can describe figurative constructions without attributing any specific linguistic marker (e.g., morpheme or syntactic feature) to the figurative meaning in the language system.

The extreme cases of collocational dependencies are the so-called idiomatically combining expressions (Croft 2001: 181): in this category, the syntactic and the semantic pattern correspond to each other, as we saw in the case of *open the eyes*. As another example, the following collocates are at the top of the list next to the verb *burst* (*into*): *flames* (11.33), *tears* (10.97), *flame* (10. 04), *giggles* (9.67). While the first and the third cases represent the primary meaning of the verb *burst*, since they refer to a sudden change of physical state, the second and the fourth collocates cannot be categorized as instances of the general construction [*burst* + CAUSED CHANGE OF STATE]. In the case of *tears* or *giggles*, the construction can be described with the correspondences *burst* → START (SUDDENLY), *tears/giggles* → EXPRESSION OF EMOTION. Finally, in the case of *bloom* (8.07) the correspondences are the following *burst* → START TO PRODUCE, *bloom* → BLOSSOMING /FLOWERING. The idiomatic combinations are thus not only independent constructions, but also cannot be assigned to a higher, more general constructional schema. Put it differently, they are

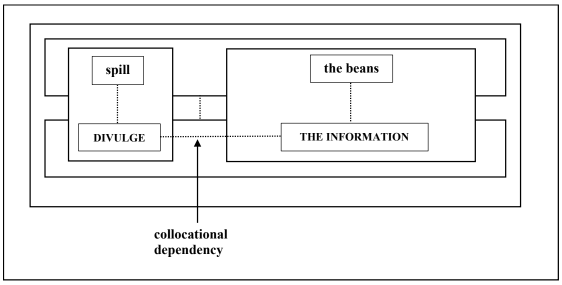

Figure 1: The schematic diagram of collocational dependencies in the idiom *spill the beans* (Croft 2001: 183)

[4] The data are from the COCA corpus (https://www.english-corpora.org/coca/), last access: 11/09/2022. The strength of collocations is measured with the MI score in the corpus.

not simply nodes on the lower level of the network of the constructicon but are nodes in themselves.

In Croft's proposal, the decisive criterion is not the presence or absence of compositionality, although it is true that – precisely because of the semantic relations symbolized by collocational dependencies – even idiomatic combinations are characterized by a degree of transparency. (Non-compositional idiomatic phrases, such as *kick the bucket*, are not transparent at all, and are therefore collocations, but not syntactically meaningful constructions – they are rather independent elements of the mental lexicon.) The crucial parameter is genericity, i.e., whether a structure can be subordinated to a higher-order construction. Collocations help to explore constructions of different degrees of abstraction along this aspect.

3.2 Valency constructions

In Herbst's constructional analysis proposal (Herbst 2014) based on valence theory, the term *collocation* does not occur, but he takes such formal patterns as a starting point for the constructional analysis that are element-specific (i.e., argument structure constructions (ASCs) are organized around specific verbs), may have a fixed word order pattern (which can be mapped with n-grams), and are based on the fact that verbs as valency carriers can open up argument positions (valency slots) in the course of construing a sentence. These initial patterns are called valency constructions in this approach which contain the potential valencies arising from the usage of a given verb and all its possible forms. As an example, two different valency constructions of the verb *give* are as follows:[5]

(2a) [SCU: NP “GIVER”]_give$_{\text{act}}$_[PCU1: NP “GIVEE”]_[PCU2: NP “ITEM GIVEN”] || Sem
(2b) And now you want to give them reputation bonus?

(2c) [SCU: NP “GIVER”]_give$_{\text{act}}$_[PCU1: NP “ITEM GIVEN”]_[PCU2: NP “GIVEE”] || Sem
(2d) they had to give it to a different teacher to be used for a different purpose

Herbst proposes not to synthesize the different valency patterns with optional constructions (e.g., implicit but expressible object arguments in the context of the verb *read*), but to treat the presence and absence of valency as different instances of valency constructions.

From valency constructions, we can generalize form, meaning, or form-meaning structures. In the first case, we obtain valence patterns that describe the context of the valency carrier with formal labels (e.g., NP, to INF in English). In the second case, we obtain participant patterns that characterize the participant roles of the event marked by the verb in a more general way (e.g., agent, patient, benrec, i.e. beneficient/recipient). At the same time, participant patterns are abstractions that can be realized by several different valency patterns. In other words, they do not prescribe the occurrence of arguments in the context of the verb. Finally, in the dimension of form-meaning pairs, the observation of concrete valency constructions arrives at general valency constructions, i.e., ASCs.

Herbst also maintains the two-step method, in which first the specific valency patterns are explored by observing the occurrences of word combinations in the corpus, and in a further step the more general ASCs can be identified, which are allostructions of the specific valency constructions at the same time. Although this approach does not give a general answer to the question of how valency constructions can be assigned to general ASCs, it takes the participant pattern (i.e., semantic motivation) as a guiding principle: all valency patterns that realize the same participant pattern can be considered allostructions of a construction. This brings us to the level of the constructeme, which is the set of a given participant pattern and all the valency constructions realizing this pattern.

The valency-based approach has not yet received a monographic explanation; thus, the applications of the analytical framework may lead to further questions. However, it seems to be a promising initiative for a data-driven description of constructions because it essentially gives priority to observable valency constructions in the description. This is also shown by the fact that Herbst while adopting the semantic coherence principle of Goldberg, complements it with the so-called valence realization principle: according to it, if the valency construction of a verb is fused with a general argument structure construction, and its

[5] The data are from the COCA corpus (https://www.english-corpora.org/coca/), last access: 01/25/2023.

participant roles are constructed as arguments, then the formal realization of the ASC must coincide with the pattern of the valency construction. This ensures that the language user's constructional knowledge does not only cover the higher-order, more general representations but element-specific constraints, i.e., lower-level patterns, are also reflected in it. Overall, Herbst considers the description of argument constructions and valency constructions as complementary steps: he calls his theory an empirical valency-based approach to argument constructions.

3.3 Patterns

Hunston's proposal (Hunston 2014) does not use the category of collocations again, but it is akin to previous approaches in that its central concept, the pattern, which is a re-occurring linguistic context around a core word, characterized by grammatical devices (e.g., dependency relations), must be identified in a rigorous corpus-driven way. No prior interpretation or grammatical theory can be assumed in the analysis until the pattern (and its semantic groups) has been identified. “Patterns, then, are a way of describing the common grammatical environment of different words and, building on these descriptions, identifying the co-occurrence of pattern and meaning. They are intentionally naïve in that they do not presuppose any particular way of interpreting word-pattern combinations” (Hunston 2014: 106).

Pattern grammar grew out of the annotation process of the Collins COUBILD English Dictionary and is thus based on the Bank of English corpus. The re-occurring grammatical context associated with each word was coded by the annotators along the lines of part of speech category, clause type and grammatical elements (e.g., prepositions) occurring in the structure. This endeavor produced a word-centered repository of patterns in English that includes also word combinations from a semantic point of view (see also Hunston and Francis 2000). Thus, the enterprise did not originally develop within the framework of collocation analysis, nor was it originally a branch of construction grammar.

Yet patterns integrate the notions of collocation (repeated co-occurrences) and colligation (grammatical choices specific to a phrase) since they contain both specific collocates and components characterized by a lexical category, the order of which is fixed. (It is no coincidence that Hunston (2014: 99) considers both Sinclairian notions as precursors to her proposal.) Thus, further analysis of the identified patterns is open to various semantic interpretations, among which Hunston highlights valency theory and frame semantics. Indeed, the patterns can be understood as element-specific valency constructions, although pattern grammar does not rely on valency theory as a theoretical background.

Hunston (2014: 112-115) emphasizes that pattern grammar is akin to construction grammar in many ways, and it can be harmonized with cognitive grammar as well. The similarities include (i) the rejection of the syntax/lexicon dichotomy, (ii) the acceptance of a tight relation of form and function, (iii) the construction-based/pattern-based conception of meaning (i.e., the rejection of word-centered meaning description), (iv) a preference for the word form over the lemma (favoring element-specific patterns over higher-level generalizations), and finally (v) a rejection of grammatical rules as abstract representations (instead, rules are redefined as generalizations of frequently reoccurring structures). Consequently, the analysis of patterns can be integrated into the cognitive constructionist approaches from a linguistic theoretical point of view.

However, patterns themselves are not constructions. While there is a large overlap between the two categories, not all constructions are patterns. For example, inversion, which is not related to specific words but rather to a group of words, such as auxiliaries, is not a specific pattern, but a general construction. Moreover, patterns are not mental representations but rather observable and identifiable usage tendencies in the corpus. By way of explanation, Hunston explicitly rejects any mentalization in modelling, although she leaves open the possibility of further interpretation of patterns. It is no coincidence that she does not regard pattern grammar as a theory of grammar, but rather as a way of describing language: “[p]ut another way, pattern grammar is not an incomplete constructional grammar, but a part of a description built on units of meaning. Pattern identification establishes order in the mass of data, but does not propose a set of mental constructs” (Hunston 2014: 115). Patterns, like collocations or valency constructions, seem to be thus the “lobby” of construction description: pattern extraction constitutes the first step of construction identification, minimizing the role of introspection on the construction grammatical approaches and maximizing the involvement of corpus data in linguistic research.

3.4 Discussion

As a modest summary, three lessons can be drawn from the overview. First, all of them instantiate methodological unidirectionality: in a bottom-up approach, these proposals start with raw data and observation, and the generalization from them towards higher-level constructions is tightly controlled. Due to this methodological commitment, a corpus-driven construction analysis can find a way out of theoretical circularity and results in not a heuristic but rather an empirically grounded interpretation of the notion of construction. The weakness of this approach is, however, that a large-scale description of the constructional network of a language is really time-consuming and needs a vast amount of effort since it begins with the exploration of corpus patterns (collocations, valency constructions or simply patterns).

Second, the presented frameworks make it possible to decrease the fluidity of the notion of construction while maintaining its flexibility. Based on corpus pattern analysis we can arrive at pure semantic generalizations (e.g., selectional restrictions), more or less schematic grammatical structures (e.g., valency carriers and their syntactic context), figurative expressions (e.g., idiomatically combining expressions) or the family of higher-level constructions (i.e., the constructeme). Put differently, the analyst can map a larger section of the constructicon without relying on their own intuition. However, the process of analyzing corpus patterns as more abstract grammatical and/or semantic configurations remains theory-driven, which means that the researcher has to make a decision what kind of theoretical perspective they will adopt, what grammatical theories (e.g., dependencies, valencies or the cognitive grammatical modelling of construal) are preferred by them. Thus, a pure empirical investigation of constructions does not seem to be achievable; nevertheless, a solid methodological foundation may serve as a vantage point for further theoretical decisions and considerations.

Third, and maybe the most important for the NLP community from the whole issue: pattern extraction is the point where NLP tools provide invaluable support to CxGs. Data are messy and do not match necessary with our expectations; but if we turn first to patterns and then form theoretical proposals about potential constructions, it may increase the reliability of our research without closing the door to discover new phenomena of language use. Moreover, it can speed up the process of analysis since the sooner we face raw data the better the precision and the recall of our analysis will be. An automatized pattern extraction process designed and tested in accordance with the demands of CxG research may also provide a remedy to the problem of a large-scale but bottom-up exploration of constructions.

4 Conclusion

This meta-theoretical and methodological study attempted to reflect on the interpretation of the concept of construction from a corpus-driven perspective. The main question of the study was how verb argument constructions can be identified in corpus analysis, and which expressions can be said to be (potential) constructions. Closely related to this is the question of whether there is a data type in corpus linguistics that can be equated with the broad notion of construction.

If the reader considers my attempt successful, they will probably agree with the following two more general conclusions. First, the notion of construction can be used in empirical research neither without reflection nor on the basis of some theoretical consensus. In a corpus-driven approach, the researcher does not rely on a pre-given model of the phenomenon under investigation but arrives at a definition and description of the phenomenon after observing and processing the data. This does not mean, of course, that we should not be aware of the diversity of linguistic constructions. It is, however, suggested that for any given construction under investigation, attributing the label of construction to a set of linguistic phenomena should not be the starting point but the end point (or at least the intermediate result) of an analysis.

Secondly, the corpus does not provide the constructions directly, therefore, a procedure needs to be developed to move from the raw data of the corpus to the constructions. Collocations can be interpreted as dependency relations with varying degrees of symbolization, valency patterns, or recurrent and grammatically more or less transparent patterns. Their precise analysis can lead us to the identification of more generic form-meaning pairings. Whichever proposal is adopted (or even if we develop our own analytical approach), the corpus contains only patterned verb–word combinations, so we should think in two steps: first, by exploring these combinations to identify constructions, and then, by further methods (e.g., by collostructional analysis), to perform a comprehensive description of the identified constructions. These two steps need not necessarily follow each other, but it is still important not to assume a priori constructions in a corpus-driven analysis.

Construction grammar and corpus linguistics can therefore be integrated in a number of ways, and we need large-scale investigation to decide which way of them will be the most appropriate. The integration is by no means pre-given, however, by achieving it we will have a better understanding of the organization of the construction.

Acknowledgments

The research was supported by the NKFI project K 129040, "Verbal Constructions of Hungarian".

References

Adele E. Goldberg. 2006. Constructions at Work: The Nature of Generalization in Language. Oxford University Press. Oxford.

Adele E. Goldberg. 2019. Explain Me This: Creativity, Competition, and the Partial Productivity of Construction. Princeton University Press, Princeton, Oxford.

Geoff Barnbrook, Oliver Mason and Ramesh Krishnamurthy. 2013. Collocation: Applications and Implications. Palgrave Macmillan, Houndmills, New York.

Joan L. Bybee. 2013. Usage-based Theory and Exemplra Representations of Constructions. In The Oxford Handbook of Construction Grammar. Eds. Thomas Hoffmann and Graeme Trousdale. Oxford University Press, New York. 49–69.

Natalia Levshina. 2015. How to do Linguistics with R. Data exploration and statistical analysis. John Benjamins, Amsterdam, Philadelphia.

Stefan Evert. 2009. Corpora and collocations. In Corpus Linguistics. An International Handbook. HSK. 29.2. Eds. Anke Lüdeling and Merja Kytö. Walter de Gruyter, Berlin, New York. 1212–1248.

Stefan H. Gries. 2013. Data in Construction Grammar. In The Oxford Handbook of Construction Grammar. Eds. Thomas Hoffmann and Graeme Trousdale. Oxford University Press, New York. 93–109.

Susan Hunston. 2014. Pattern grammar in context. In Constructions Collocations Patterns. Eds. Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber. De Gruyter Mouton, Berlin, Boston. 99–120.

Susan Hunston and Gill Francis. 2000. Pattern Grammar: A corpus-driven approach to the lexical grammar of English. John Benjamins, Amsterdam, Philadelphia.

Thomas Herbst. 2014. The valency approach to argument structure constructions. In Constructions Collocations Patterns. Eds. Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber. De Gruyter Mouton, Berlin, Boston. 167–216.

Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber. 2014. From collocations and patterns to constructions – an introduction. In Constructions Collocations Patterns. Eds. Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber. De Gruyter Mouton, Berlin, Boston.1–8.

Violeta Seretan. 2011. Syntax-Based Collocation Extraction. Springer, Dordrecht.

William Croft. 2001. Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press, Oxford.