A Lattice Based Algebraic Model for Verb Centered Constructions

Bálint Sass(✉)

Research Institute for Linguistics, Hungarian Academy of Sciences,
Budapest, Hungary
sass.balint@nytud.mta.hu

Abstract. In this paper we present a new, abstract, mathematical model for verb centered constructions (VCCs). After defining the concept of VCC we introduce proper VCCs which are roughly the ones to be included in dictionaries. First, we build a simple model for one VCC utilizing lattice theory, and then a more complex model for all the VCCs of a whole corpus combining representations of single VCCs in a certain way. We hope that this model will stimulate a new way of thinking about VCCs and will also be a solid foundation for developing new algorithms handling them.

Keywords: Verb centered construction · Proper VCC · Double cube
Corpus lattice

1 Verb Centered Constructions

What is a *verb centered construction (VCC)*? We will use this term for a broad class of expressions which have a verb in the center. In addition to the verb, a VCC consists of some (zero or more) other linguistic elements which are or can occur around the verb. In this paper, the latter will be PP and NP dependents of the verb, including the subject as well. The definition is rather permissive because our aim is to cover as many types of VCCs as we can, and provide a unified framework for them.

Sayings (*the ball is in your court*) meet this definition just as verbal idioms (*sweep under the rug*), compound verbs/complex predicates (*take a nap*), prepositional phrasal verbs (*believe in*) or simple transitive (*see*) or even intransitive verbs (*happen*). The first example above shows that it is useful to include the subject, as the concrete subject can be an inherent part of a VCC.

As elements of a VCC, we introduce the notion of *bottom*, *place* and *filler*. The bottom is the verb, there are places for PP/NP dependents around the verb, and fillers are words which occur at these places. Using this terminology, in *sweep under the rug* there is a place marked by the preposition *under* and filled by the word *rug*. Similarly, in *take a nap* there is a place for the object (designated

The original version of this chapter was revised: An acknowledgement has been added. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-00794-2_57

P. Sojka et al. (Eds.): TSD 2018, LNAI 11107, pp. 231–238, 2018.
https://doi.org/10.1007/978-3-030-00794-2_25

by word order in English) filled by *nap*. The VCC *believe in* demonstrates the notion of a *free place*, marked by the preposition *in* and not filled by anything. We can talk about different classes of VCCs – fully free, partly free, fully filled – according to how many places and fillers they have. *Take part in* has one filled place (object) and one free place (*in*), showing that a VCC can be a compound verb and a prepositional phrasal verb at the same time.

Have a closer look at *sweep under the rug*. We find that this VCC is in a certain sense not complete. In fact, it should have two additional (free) places: one for the subject and another for the object. Let us use the following notation for VCCs: [sweep + subj + obj + under ↶ rug]. First element is the verb, places are attached by +, and fillers are attached to the corresponding place by ↶. (This representation does not indicate word order: places are taken as a set.) If we narrowed down our focus to a certain kind of VCCs, we would obtain expressions which are incomplete in the above sense. For example, in their classical paper Evert [5] search for proposition+noun+verb triplets, they obtain for example *zur Verfügung stellen* which is clearly incomplete: it lacks free subject and object places.

Table 1. Illustrating the notion of *proper VCCs*. Clearly, the proper VCC is transitive *read* in the first sentence and *take part in* in the second (together with the free subject place). Other VCCs of these sentences are evidently not proper.

John reads the book.	
VCC	Proper?
[read + subj + obj]	+
[read + subj + obj ↶ book]	–
John takes part in the conversation.	
VCC	Proper?
[take + subj + obj]	–
[take + subj + obj ↶ part]	–
[take + subj + obj ↶ part + in]	+
[take + subj + obj ↶ part + in ↶ conv.]	–

As we see, not all VCCs are multi-word, but most of them are multi-unit at least. Whether a certain VCC is multi-word or not depends on the language: the counterpart of a separate word (e.g. a preposition) in a language can be an affix (e.g. a case marker) in an other. Our target is the whole class of VCCs, so we do not lay down a requirement that a VCC must be multi-word.

All sentences (which contains a verb) contain several VCCs which are substructures of each other. From these, one VCC is of special importance: the *proper VCC*. This notion is essential for the following. The proper VCC is complete, that means it contains all necessary elements, and clean, that means it does not contain any unnecessary element. It contains free places constituting complements and does contain free places constituting adjuncts. It contains fillers which are idiomatic (or at least institutionalized [8]) and does not contain fillers

Table 2. A Hungarian example for interfering places in VCCs. Places follow their fillers in this table because places are cases in Hungarian: *-t* is a case marker for object, and *-rA* for something like English preposition *onto*.

Verb	Filler	Place	Filler	Place
vet	*pillantás*	*-t*		*-rA*
cast	glance	obj		onto
Cast a glance onto something = look at something				
vet		*-t*	*szem*	*-rA*
cast		obj	eye	onto
Cast something onto (somebody's) eye = reproach somebody for something				

which are compositional. In fact, we look for a combination of elements beside the verb that, together with the verb, form a *unit of meaning* [10]. In short, the proper VCC is the verbal expression from a given sentence which can be included in a dictionary as an entry (Table 1).

Notice that we have one certain set of linguistic tools for expressing places of VCCs in a language: word order and prepositions in English, prepositions and case markers in German, postpositions and case markers in Hungarian etc. Since we use them both for free and filled places, they can interfere with each other beside a verb: place `A` can be free and place `B` filled beside verb `V` in a VCC, while place `B` can be free and place `A` filled beside the same verb in another (Table 2).

We present an algebraic model for VCCs in the next sections.

2 Model for One VCC: The Double Cube Lattice

Let us take a cube, and generalize it for n dimensions. The 1-dimensional cube is a line segment. The 2-dimensional cube is a square. The 3-dimensional cube is the usual cube. The 4-dimensional cube is the tesseract. Now, let us create the so called *double cube* by adding another cube in every dimension to make a larger cube whose side is twice as long. The 1- and 2-dimensional double cube can be seen in Fig. 1, the 3-dimensional double cube can be seen in Fig. 2. The n-dimensional double cube consists of 2^n pieces of n-dimensional cubes.

Double cubes should always be depicted with one vertex at the bottom and one at the top. Edges of double cubes are directed towards the top. Notice that supplemented with these properties double coubes are in fact *bounded lattices* [7, part 11.2], and the ordering of the lattice is defined by the directed edges. Epstein [4] calls these structures n-dimensional Post lattices of order 3. Post lattices are a generalization of Boolean lattices ('simple cubes') as Boolean lattices are the same as Post lattices of order 2.

Now, we relate VCCs to these kind of general structures. A double cube will represent the fully filled VCC of a verbal clause taken from a corpus together

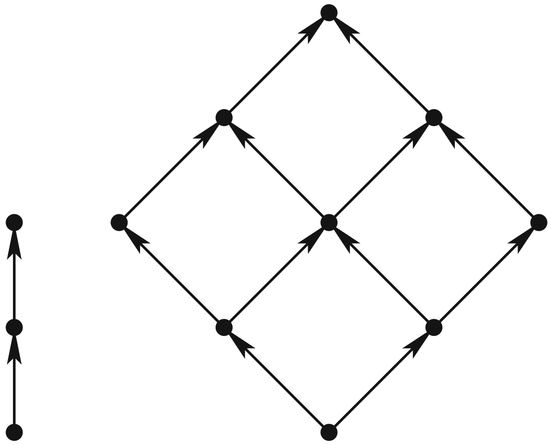

Fig. 1. The 1-dimensional double cube which consists of two line segments, and the 2-dimensional double cube which consists of four squares.

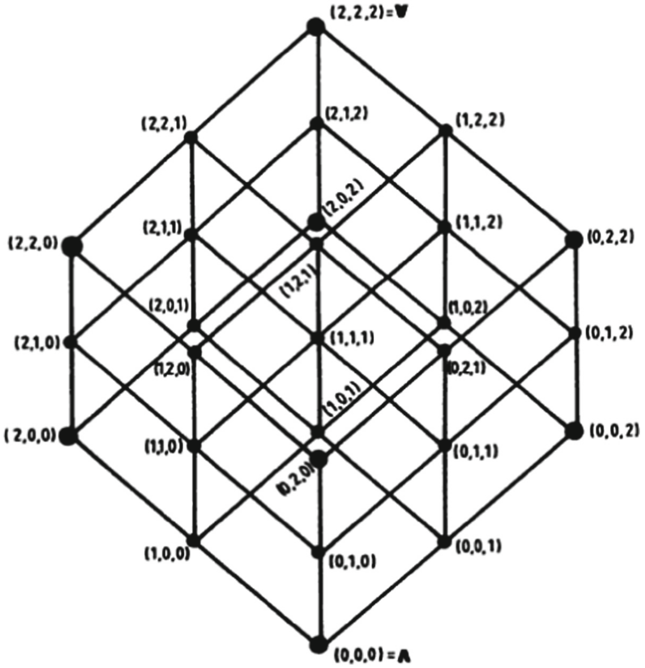

Fig. 2. The 3-dimensional double cube made from 8 usual cubes. This figure is taken from [4, p. 104] or [3, p. 309].

with all of its sub-VCCs. The dimension of the double cube equals with how many places are there in the fully filled VCC. All vertices are sub-VCCs of the fully filled VCC, while edges are *VCC building operations*. There are two such operations: *place addition* (represented by +) and *place filling* (represented by $\curvearrowleft$). The model is based on the idea that places and fillers are both kinds of elements, so place addition and place filling are treated alike as VCC building operations working with elements.

Of course, place addition must precede place filling with respect to a specific place. This very property is what determines the cubic form of the lattice. The bottom of the lattice is the bare verb, the top of the lattice is the fully filled VCC: it contains all fillers which present in the clause regardless whether they are part of the proper VCC or not. The proper VCC itself is one of the vertices (cf. Table 1). Figure 3 shows the first sentence of Table 1 as an example. Representing the second sentence is left to the reader. This would require a 3-dimensional double cube and the proper VCC would be the vertex marked by (1, 2, 1) in Fig. 2 if we define the order of places according to Table 1.

Our approach follows the traditional theory of valency [9] in some aspects, as we talk about slots beside the verb and take the subject as a complement as well, but we deviate from it in other aspects, as we do not care what kind of complements a verb can have in theory, but take all dependents we find in the

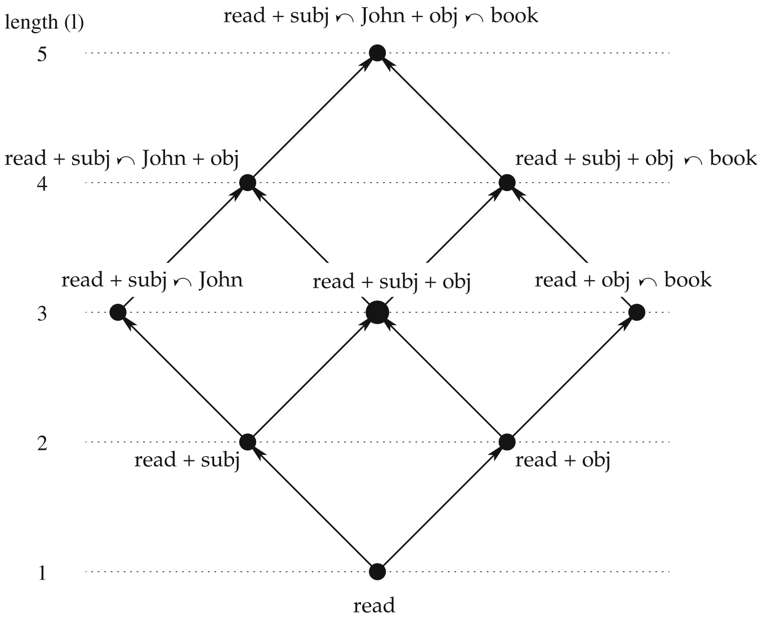

Fig. 3. The double cube representation of the clause *John reads the book*. The fully filled VCC is at the top of the lattice. The proper VCC is [read + subj + obj] which can be seen in the center marked with a larger dot. Length of a VCC (see left side) is defined as how many elements it consists of.

corpus instead, dealing complements and adjuncts in a uniform way, following the full valency approach [1] essentially.

As we see, the double cube serves two purposes at the same time: on the one hand, it is a representation of a verbal clause, on the other hand, one vertex of the double cube marks the proper VCC of this clause.

It is important to see that a double cube is not simply a graphical form of a power set. Unlike the power set where two fundamental possibilities exist (namely being or not being an element of a set), we have three possibilities here concerning a place of a VCC: the place does not exist, the place exists and it is free, the place exists and is filled by a given filler. Obviously, it is important to discriminate between no object (*happen*), free object (*see*) and filled object (*take part*). Graphical form of a power set would be a Boolean lattice, Post lattices are a generalization of them (from 2 to 3) as we mentioned earlier. The mere fact that a place occurs in a clause does not mean that this place will be a part of the proper VCC of the clause. The model must give some opportunity to omit certain places from the original clauses if necessary. The double cube model meets this requirement appropriately.

Some grammatically incorrect expressions can be noticed in Fig. 3 (e.g. *read+obj*). As the double cube is a formal decomposition of the original clause, it is not a problem to have some vertices representing ungrammatical expressions, the only thing which should be ensured that the chosen proper VCC is grammatically correct at the end.

3 Model for a Whole Corpus: The Corpus Lattice

Using lattice structures defined above, a complex model can be built which represents all VCCs occurring in a corpus. So far, we have built double cubes from elements, now the double cubes themselves will be the building blocks for assembling the corpus lattice. Having double cubes introduced explicitly is what allows us to build this larger lattice. The corpus lattice is considered as one of the main contributions of this paper. As it represents the distribution of all free and filled places beside verbs, we think that it is a representation which can be the basis for discovering typical proper VCCs of the corpus.

We define the *lattice combination* ($\oplus$) operation for lattices having the same bottom. Let $L_1 \oplus L_2 = K$ so that K a minimal $\wedge$-semilattice which is correctly labeled and into which both lattices can be embedded. In other words: let it be that $L_1 \subseteq K$ (with correct labels) and $L_2 \subseteq K$ (with correct labels) and K has the minimum number of vertices and edges, and labeled edges of $L_1 \cup L_2$ occurs only once in K (Fig. 4).

We build the corpus lattice this way: we go through the corpus, take the verbal clauses one by one, and combine the double cube of the actual clause to the corpus lattice being prepared using the $\oplus$ operation defined above. (As only lattices having the same bottom can be combined, we will obtain a separate $\wedge$-semilattice for every verb. One such $\wedge$-semilattice and the set of all both can be called corpus lattice).

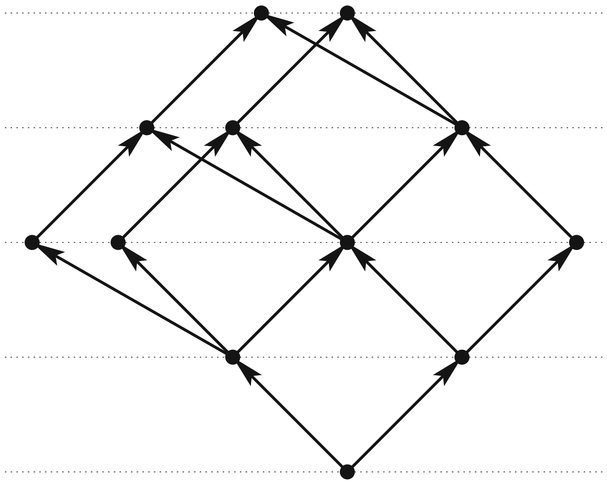

Fig. 4. An illustration of the lattice combination operation. This structure is the corpus lattice representation of a small example corpus containing only two sentences: *John reads the book* and *Mary reads the book*. It is a ∧-semilattice: the bottom (the verb) is unique, the top is clearly not.

Remark: in category theory, the lattice combination operation corresponds to the *coproduct* [6, pp. 62–63] defined on the lattice of corpus lattices in which the ordering is defined by the above embedding.

If we compare our model to other approaches of verbal relations (e.g. verb subcategorization, TAG or FrameNet), main difference can be phrased as follows. Firstly, our model puts great emphasis on filled places, and accordingly on complex proper VCCs (which have filled places and possibly free places as well), connecting our approach to multiword expression processing [2]. Secondly, the aim of our model is to represent not just one VCC but all VCCs of a corpus together including their relationships to each other, in order to be able to tackle proper VCCs based on this combined model. The corpus lattice is the tool which realizes this aim projecting VCCs onto each other in a sense, the double cubes can be considered as an aid for creating the corpus lattice.

4 Summary and Future Work

In this paper, we presented a model for VCCs. Hopefully, this model will allow us to talk about this type of constructions in a new way and it will also be a suitable basis for developing algorithms handling them. The model provides a unified representation for all kinds of VCCs being multi-word or not, regardless of the language they are in, and also regardless whether they have free or filled places opening up an opportunity to solve the interference problem exemplified in Table 2.

Our main future aim is to discover proper VCCs. To achieve this, new methods are needed which collect all the required places and determine whether a place is free or filled by a certain filler, in order to make the VCCs complete and clean. The corpus lattice – equipped with corpus frequencies at every vertex – is an appropriate starting point for developing this kind of new algorithms. We think that proper VCCs are at some kind of thickening points of the corpus lattice. The prospective algorithm would move through the corpus lattice (top-down or bottom-up) vertex by vertex, until it reach proper VCCs at such points. For this type of algorithms it is needed to be able to effectively advance from one vertex to another differing only in one element. Our representation is suitable exactly for this purpose.

Another future direction can be discovering parallel proper VCCs. They may be useful for tasks where multiple languages are involved (e.g. machine translation). On the one hand, proper VCCs are not to be translated element by element, they need to be known and interpreted as one unit. On the other hand, being complete, they can be corresponded to each other if not element by element, then at least free place by free place. For example Dutch *nemen deel aan* and French *participer à* have completely different structure (multi-word vs. single-word), but sharing the same meaning they both have one free place (beside the subject). We think that parallel proper VCCs can be discovered applying our model to parallel corpora in a way.

Acknowledgement. This research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (case number: BO/00064/17/1; duration: 2017-2020).

References

1. Čech, R., Pajas, P., Mačutek, J.: Full valency. Verb valency without distinguishing complements and adjuncts. J. Quant. Linguist. **17**(4), 291–302 (2010)
2. Constant, M., et al.: Multiword expression processing: a survey. Comput. Linguist. **43**(4), 837–892 (2017)
3. Epstein, G.: The lattice theory of Post algebras. Trans. Am. Math. Soc. **95**(2), 300–317 (1960)
4. Epstein, G.: Multiple-valued Logic Design: An Introduction. IOP Publishing, Bristol (1993)
5. Evert, S., Krenn, B.: Methods for the qualitative evaluation of lexical association measures. In: Proceedings of the 39th Meeting of the Association for Computational Linguistics, pp. 188–195. Toulouse, France (2001)
6. Mac Lane, S.: Categories for the Working Mathematician. GTM, vol. 5, 2nd edn. Springer, New York (1978). https://doi.org/10.1007/978-1-4757-4721-8
7. Partee, B.H., Ter Meulen, A., Wall, R.E.: Mathematical Methods in Linguistics. Kluwer Academic Publishers, Dordrecht (1990)
8. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: a pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45715-1_1
9. Tesnière, L.: Elements of Structural Syntax. John Benjamins, Amsterdam (2015)
10. Teubert, W.: My version of corpus linguistics. Int. J. Corpus Linguist. **10**(1), 1–13 (2005)