Dictionary of Verbal Contexts for the Romanian Language

Ana-Maria Barbu
Institute of Linguistics, Romanian Academy, Bucharest
E-mail: anamaria.barbu@g.unibuc.ro

Abstract

This paper presents a dictionary of verbal contexts for Romanian, which comprises 600 verbs and over 2,000 meanings with one or more valency patterns. It is manually built but based on corpus information, and is developed both for teaching Romanian to foreigners, by its printed version, and for computational linguistics, by its XML format and consistent principles and conventions of the design. The dictionary is rich in information, including lexical, grammatical and semantic features of the complements, morphosyntactic variants occupying an argument position, dependencies between complements induced by control, raising and predication phenomena and verbal alternations, as variants of valency patterns with the same meaning. The paper offers details about all this information, the building procedure and some problems that needed to be solved during our work. This enterprise is far from being finished, because further work has to be done to improve the actual encoding and add new types of information, such as semantic roles or diathesis uses, for growing the number of entries and for getting different kinds of generalizations.

Keywords: verb pattern, verbal context, valency, argument, complement, Romanian verbs

1 Introduction

This study presents a work focused on the concept of verb valency, seen as a configuration of dependents (i.e. arguments or complements) lexically required by a verb (or a predicate, in general). The concept of valency, as is well known, is rooted in Tesnière's theory of dependency grammar, however, as Herbst (1999: 2) points out, the beginnings of a valency theory as such are not found in works within dependency grammar, but within foreign language teaching; and they are marked by the first valency dictionaries, such as Helbig and Schenkel (1968) or Engel and Schumacher (1976). Since then, both valency theory and valency dictionaries have shown their scientific potential and utility. Nowadays, important efforts are being made to build such dictionaries in more and more languages, especially with the support of computational and corpus linguistics, which, in turn, is a main beneficiary of such a valuable resource.

In this vein, the *Dictionary of Verbal Contexts for Romanian Language* (hereafter DVCRL) we present here represents a similar endeavor. This dictionary is manually built but mainly based on internet data, and comprises 600 verbs and over 2,000 meanings with one or more verb patterns (or complementation patterns), both in machine-readable format and in printed version (see Barbu 2017). This work started about ten years ago and has, as models of that time, FrameNet (Johnson & Fillmore 2000), VerbNet (Kiper et al. 2000), CPA (Hanks & Pustejovsky 2005, Hanks 2006) and VALLEX 2.0 (Žabokrtský & Lopatkova 2007), as described and analyzed in Barbu (2008). The dictionary building had two phases. The first was developed by five linguists, within a national project spanning three years, which collected about 3,000 verb entries.[1] In this phase, the description of valencies was limited mainly to obligatory complements (minimal valencies) and to the verb meanings in the Romanian Explanatory Dictionary (DEX 1998). The lack of further funding and the ned to solve the

1 The first version of the dictionary was created within the CNCSIS project nr. 1156/2005, during the years 2005-2007.

shortcomings observed in the previous version urged us to take over the improvement of the dictionary on our own, being aware of how important is to obtain such a linguistic resource, and that such work has to be done as well as possible. In these conditions, the enterprise advanced very slowly and could only cover much fewer entries, due to the fact that it was supported by only one linguist, who for a while was working outside his job duties.

The improvements made in the second phase mainly address the following aspects. On the one hand, we have paid special attention to facultative or optional complements[2] and adjunct-like elements (or modifiers), also encoded in FrameNet, Vallex and CPA. For instance, any directed-movement verb has a Path, including Origin and Goal, as its lexically-governed complements, but they are facultative because can be omitted depending on the communication context, like Path and Goal in *John has gone from the school*. Furthermore, a verb like *a căuta* 'search' is very frequently accompanied by a Locative modifier, unlike other types of modifiers (e.g. Time or Manner), which justifies the inclusion of a Locative element in the complementation pattern of this verb. Due to these facts, we preferred the term *verbal context*, instead of *verb valency*. While *verb valency* (predicate-argument structure or subcategorisation frame) generally refers to *minimal* number of arguments required by a verb for accomplishing its meaning, *verbal context* includes, as understood here, adjunct-like elements which are frequently used in events centered on a certain verb, in the same line as Fillmore's *semantic frame*.

On the other hand, we have consolidated the concept of *argument position,* allowing morpho-syntactic variants (Barbu & Ionescu 2007:45), in the same vein as that mentioned later in (Przepiórkowski et al. 2014: 2786), as we show below. Another improvement refers to the specification of alternations, considered in a more restricted sense than Levin's (1993), that is, as sets of verb patterns with, roughly, the *same* meaning. The meanings themselves of each verb had to be re-thought, as we point out below in the next section. We have not left aside the information indicating dependencies between different complements of the same pattern either. Such dependencies are induced by control and raising phenomena or small clauses. It is worth mentioning that all the verb contexts are valid for declarative sentences and the active voice, and that many regular transformations or variations, which do not have a lexical nature, are not caught in the encoding of this dictionary.

In what follows we present some aspects of the building procedure used in this project, with a short explanation of why we had to dispense with DEX meanings and re-think them. In Section 3, we detail the information structure of an entry and verb pattern, while we also explain some of our encoding options. This section deals with the topics of alternations and XML representation, as well. Some challenging aspects of our enterprise are treated in Section 4, and we conclude the paper with general remarks and plans for further work.

2 The Building Procedure

Due to the fact that DVCRL is manually built, it is not founded on valency samples extracted from a large corpus, as other approaches are, but on meaning-oriented bases, by tightly correlating meanings with contexts in use. Thus for each verb meaning taken from monolingual dictionaries, we build contexts by intuition and check them on the internet (seen as a corpus, cf. Kilgarriff & Grefenstette 2003), which can bring up new, unexpected contexts.[3] The search online cannot be done by using part-of-speech tags, and therefore for obtaining rarer structures we used some lexical expressions we know are used, eventually adapting them with the wildcards allowed by Google search engine.

2 Actually, in this paper we settle a new distinction between facultative and optional complements.

3 Note that the Romanian language has had a large part-of-speech annotated corpus only since the end of 2017, and therefore we could not adopt a corpus-oriented methodology.

With this we have noticed, quite surprisingly, that a corpus, however big it could be, does not cover a good number of language facts. Despite these difficulties, in our opinion, this work method, which combines corpus analysis with native speaker intuition, has some advantages over an exclusively corpus-oriented approach, because it ensures benefits like the following: access to less frequent uses of a verb (for less used meanings); human generalization of the information unavailable in a corpus (e.g. semantic roles or selectional restrictions); and the correspondence between the valency frame and the meaning, ready to use, for instance, in computational semantic disambiguation.

In the first phase, for each verb, we adopted its meanings from the Romanian Explanatory Dictionary (DEX 1998) and assigned the minimal verb contexts for each of them, as described in Barbu and Ionescu (2007) and Barbu (2008). But this turned out to be a bad start, because we could not obtain a convenient correspondence between the patterns of a verb and its DEX meanings. For instance, for the transitive verb *a aplauda* 'applaud' DEX offers only the meaning "to express satisfaction, approve or admiration for something by clapping the hands". By a careful inspection of the verb contexts in use, one can find two embarrassing facts: 1. the subject of this verb can refer to entities without hands (e.g. institutions or organizations) so that no clapping of hands can happen – actually it shows up a new meaning: "to praise"; 2. when clapping of hands is involved, and only in this case, the direct object can be omitted. In order to capture these peculiarities in our encoding, we needed to set two different semantic restrictions: +human for "human creature (with hands, feet, eyes, etc.)" and +person for "physical or juridical person", and two different verb patterns equivalent (in Romanian) to those in example (1) (see Figure 1 below for the structure of such an entry):

(1) applaud
I.
1. NP [nom +human], 2. (fac.) NP [acc]
to clap the hands (for expressing the approval of someone or something): *The audience applauded (actors) at the end of the show in standing ovation.*//
II.
1. NP [nom +person], 2. NP [acc]
to praise: *Coalition applauds New York City's universal physical education initiative.*//

Note that in the first verb pattern (I.) the noun phrase in accusative (NP[acc]) can be omitted (it is marked by (fac.)), which is mirrored in the pattern example by the omissible word *actors* placed inside parentheses. Instead, in the second verb pattern (II.) the direct object is obligatory, otherwise odd utterances such as **Coalition applauds* are obtained.

As one can see, different semantic restrictions (e.g. +human versus +person) can trigger and justify different verb patterns. Actually, finding what particularizes a certain pattern was the most important challenge of our work.

As a general method of dictionary building, we keep DEX as a reference dictionary but we have to reconsider the verb meanings in DEX. Not every meaning in DEX also deserves to be mentioned in our dictionary, because there are some meanings with the same valency structure. At the same time, we discovered meanings not registered in DEX. The problem of this meaning mismatch has been previously touched on by Herbs et al. (2004: xxxviii), who claim: "Establishing senses according to their valency patterns in some cases results in a rather different identification of senses than in conventional dictionaries".

The inventory of the verb entries is established according to decreasing occurrence frequencies of the verbs extracted from a newspapers corpus. This led us to include, in DVCRL, verbs used in mass-media rather than in the core vocabulary of Romanian, so that verbs such as *a mânca* 'eat' or *a dormi* 'sleep' could be missing.

For editing the dictionary, we used the Professional Lexicography Software TshwaneLex (tshwanedje.com). It has many functions, but we only mention the XML editor, the possibility of RTF or HTML export and the so-called WYSIWYG ("what you see is what you get") view. These help to easily get the dictionary in machine-readable and printed formats. Even if the software is provided with DTDs (Document Type Definition schemata) appropriate for bi- and mono-lingual dictionaries, we had to get rid of them and to build a DTD specific to the information structure of verbal context entries (see §3.5). Before writing the DTD, one has to get the final form of the information structure and XML annotation schema, because it is very risky to change the DTD after beginning the work of dictionary expansion. In general, adding new elements is possible, if child relations are not changed. Therefore, we firstly worked out, in text format, a good number of verbs (with one or several meanings/patterns) in order to capture the relevant and general information structure.

3 The Information Structure

3.1 The Structure of an Entry

An entry is headed by a verb lemma and has the information structure shown in Figure 1, illustrating the verb *a concura* 'take part in a tournament' (I), 'compete' (II) or 'act together' (III). For each lemma, one or more *verbal contexts* are given, numbered with I, II, etc. Such a verbal context includes:

a) Morphological information about the verb. For instance, the context III is used only at the 3rd person (singular or plural) and it is marked by V[3], accordingly. This type of information can also refer to specific mood or tense, as well as to pronominal or negation clitics of the verb, and it is enclosed inside square brackets preceded by V (see also V[se] in Figure 1 context II, which indicates that the verb has the reflexive clitic *se* in that verb pattern).

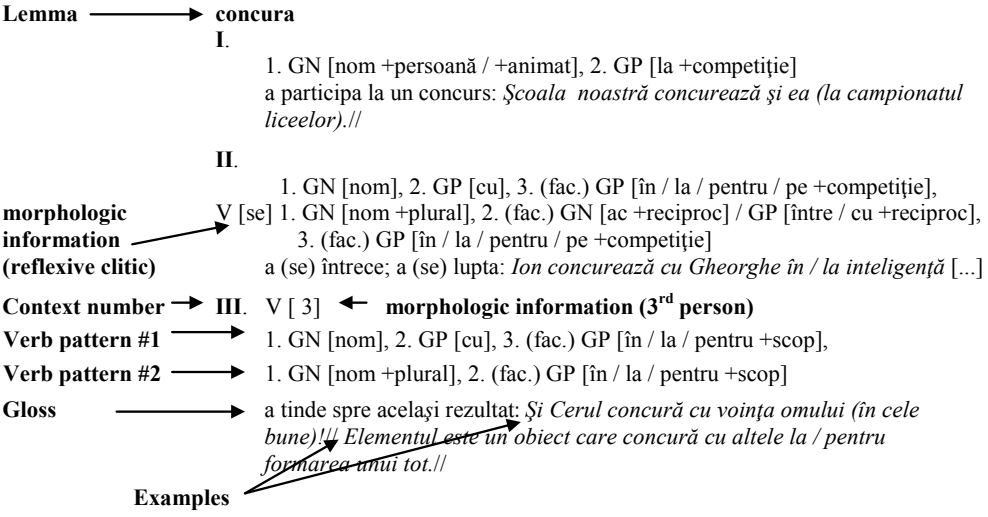

Figure 1: Entry structure.

b) One or more *verb patterns*. As Figure 1 shows, verb alternations (of two or more patterns) are described. The structure of a verb pattern is explained in sub-section 3.2. In the first phase of the project we considered that the morphological information applies to a certain context/meaning and we have assigned this information to the context level (see context III), that is higher than the pattern level. During the work, it turned out that such an information type was also needed at the pattern level in cases of alternations. For instance, context II displays the reciprocal alternation, which has a variant

implying, in Romance languages, the use of a reciprocal/reflexive clitic on the verb. This fact compelled us to make a late change to the dictionary DTD, fortunately without any perturbation of the existing annotation scheme.

c) One or more glosses by means of synonyms and sometimes paraphrases. In general, we strive to choose the synonyms that can replace the lemma verb in the given examples. If examples found in corpus fit the same verb context but require non-synonymic glosses, then several senses are provided (marked with a., b., etc.).

For instance, in the example (2), the Romanian verb *a cuprinde* 'embrace' has three meanings: a. to hold close with the arms, usually as an expression of affection: *John embraced Mary around her waist*, b. to surround or enclose: *The mist embraced the hills* and c. to include or contain as part of something broader: *The contract embraces the elements of work*.

(2) cuprinde
1. GN [nom], 2. GN [ac]
a. a îmbrățișa: *Ion a cuprins-o pe Maria de mijloc.*//
b. a închide în sine: *Ceața a cuprins dealurile.*//
c. a conține; a fi alcătuit din: *Contractul cuprinde elementele muncii prestate.*//

d) One or more examples. We give examples for each pattern alternation and each description variant marked with '/'. We propose our own examples whenever they express a common use, or are extracted from internet (and adapted by shortening or person name deletion) when they show special uses or meanings.

e) One or more idioms centered on the lemma (if it is the case). Idioms can be considered verbal contexts with lexicalized complementation, and this is the reason for including them in our description. The verb *a ajunge* 'reach' has the idiom in example (3), where what follows the '=' mark is its gloss. An idiom can have a fixed part and a mobile one, differentiated in (3) by two character types: italics and normal, respectively. The mobile part can take different references (see *cuiva* 'to somebody' in (3)) or can be inflected.

(3)	*a-i*	*ajunge*	*cuiva*	*cuțitul*	*la os*	=	a fi într-o situație disperată
	to-clitic.dat	reach	somebody.dat	knife.the	to bone	=	to be in a desperate situation

3.2 The Structure of a Verb Pattern

A verb pattern describes the complements of the verb lemma corresponding to a certain meaning and it has the information structure shown in Figure 2.[4]

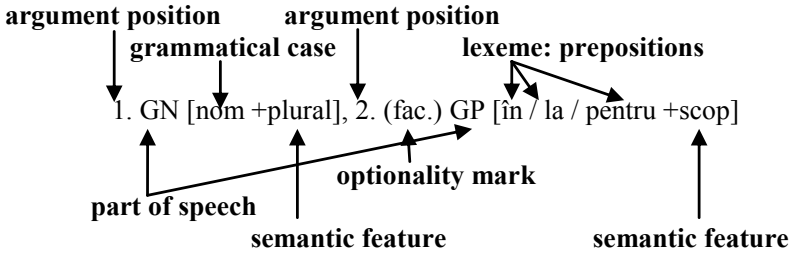

Figure 2: Verb pattern.

4 GN=noun phrase (NP), GP=preposition phrase (PP), nom=nominative case, ac=accusative case, +plural=plural, fac.= facultative, în/la/pentru=in/to/for, +scop=goal.

As one can see in Figure 2, a verb pattern contains one or more *argument positions*. We talk about *positions* instead of arguments or complements, because such a position can be occupied by different types of phrases. For instance, the predicative element of the verb *be* can be a noun phrase: *He is a doctor*, an adjective phrase: *He is very happy* and so on, and all of these types of phrases occupy the same position (the predicative one) as variants. For Romanian, a free word order language, there is no connection between the positions in the verb pattern and the complements order in utterances. However, by convention, the first position is reserved for the subject.

As is well known, a position is obligatory or not, but it is worth pointing out some further aspects of this here. By obligatory complement one usually understands a complement which cannot be excluded from the immediate context of the verb. However, many linguists, such as Allerton (1975), Fillmore (1986) or Herbst (1999), point out that there are unexpressed complements that have to be obligatorily recovered from the larger linguistic or extra-linguistic context. In other words, such a null (or recoverable) complement, even if it is not expressed, is not deleted from the argument structure of the verb. In DVCRL, we consider such a complement obligatory as well, and consequently we do not mark it as optional; however, in examples we put it in parenthesis as omissible in immediate context of the verb. At the same time, the optional complements are those which are legitimated by the meaning of the verb but can be relevant in some contexts and irrelevant in others. Such complements are marked in DVCRL with (fac.) 'facultative' (see Figure 2). More than that, we have found a special case of optional complements, when any of several complements of the same verb pattern can be omitted but at least one has to be expressed. For instance, in Romanian there are light verbs whose meanings are strictly determined by their complements. Such a verb is *a da* 'give' which acquires a motion meaning if it is accompanied by motion complements (i.e. Goal, Origin or Path). In the example (4)a the verb has all the complements expressed.

(4)

(a) El dă bagajele jos din tren pe geam.
He gives the luggage down from the train through the window.
'He throw his luggage out of the train window.'

(b) El dă bagajele jos.
He gives the luggage down.

(c) *El dă bagajele.
He gives the luggage.

In (4)b there is only one complement (the adverb *jos* 'down'), expressing the Goal, which cannot be further deleted without making the utterance ungrammatical for the intended meaning, (4)c. For these cases of optional complements, we set the mark (opt.) 'optional', obligatorily appearing in a series of two or more complements in a verb pattern and having the meaning that any complement in the series can be missing but not all of them. The difference between (fac.) and (opt.) is therefore given by the fact that in the former case the complement can be missing independently of the other complements, while in the latter case the complement can be missing only if at least one of the other complements marked by (opt.) is expressed.

The necessity of working with argument positions was actually dictated by the optional complements, which raise the question as to whether or not they can co-occur when expressed. If they do co-occur in standard configuration (that is, not involving coordination, restriction or appositive specification, for instance), then, logically, they occupy different argument positions. This turns out to be the case for obligatory complements too, when establishing the morpho-syntactic variants (i.e. phrases of different parts-of-speech) assigned to a certain argument position. In other words, the variants exclude each other, so that the complement of *quantity* (marked +quantity) of the verb *weigh*, for example, can be expressed by a noun phrase (e.g. *The bag weighs two pounds*) or an adverb phrase (e.g. *The*

bag weighs heavily) but not by both at the same time (e.g. **The bag weighs two pounds heavily*). Such a verb should have the verb pattern in example 5, where '/' separates variants:

(5) 1. NP[nom +object], 2. NP[nom +quantity] /AdvP[+quantity]

The concept of argument position with morpho-syntactic variants is also encountered in Przepiórkowski et al. (2014: 2786), but the test used there to detect the occupants of the same position is not the co-occurrence principle but the coordination. In our opinion, coordination is too sensitive to all kind of linguistic and extra-linguistic factors, so that a coordination failure could lead to incorrect results. For instance, in Romanian an example such as *Sacoșa cântărește două livre și greu* 'The bag weighs two pounds and heavily' is very odd. By applying the coordination principle the two phrases should occupy different positions, and that is wrong.

A position can be fulfilled by one phrase (or more) (see Figure 2). A phrase is characterized by its part-of-speech, grammatical or lexical information and semantic information. Grammatical information refers to grammatical cases for noun and adjective phrases (see *nom* for nominative in Figure 2), or refers to mood for verb phrases (i.e. participle, gerund or supine). Lexical information is offered for governed preposition and adverb phrases (see the prepositions series in Figure 2), and it specifies conjunctions or relative adverbs in subordinating clauses. Semantic information (preceded by +) mainly represents semantic (or selectional) restrictions, but it can reach the generality of a semantic role such as Goal (see +*scop* in Figure 2) or Source, Cause etc., in order to characterize, for instance, a series of possible prepositions with appropriate meaning without specifying them explicitly. Note that this dictionary does not use Thematic Roles nor micro-roles (cf. Hartmann et al. 2013), nor even grammatical functions. We focus on "visible", uncontroversial marks such as grammatical cases (note that Romanian is a fully inflectional language), which express, in general, certain Thematic Roles and grammatical functions (e.g. accusative is Patient/Theme and direct object).

3.3 Alternations

As pointed out in Kettnerová et al. (2012: 434), despite the growing attention paid to alternations in theoretical linguistics (especially since Levin's seminal work (1993)), there are few valency lexica approaching this topic. DVCRL includes alternations explicitly and consistently. However, in contrast with Kettnerová et al., we have completely ignored what the authors call "grammaticalized alternations", more precisely the regular transformations of diatheses and other Romanian linguistic phenomena, such as possessive dative or object duplication. Even among the morpho-syntactic variations on the argument positions, which Levin counts as alternations too (see, for example, 1993:43 Preposition drop alternation), there are some *regular* phenomena such as replacing an NP with a free-relative clause: *My friend believes me* can be replaced by *Who is my friend believes me* or a locative, time or manner complement, by its corresponding relative clause: *I go there*, by *I go where I like*. These kinds of replacements apply for any verb and any argument position, so it would be redundant to mention them in the dictionary.

Instead, we have encoded what Kettnerová et al. call "lexical alternations", to which we have added the reciprocity transformation, which is word-oriented, despite its regular transformation pattern. In Figure 1, both verbal contexts II and III display reciprocity alternation, but of different types. While the first requires a reflexive clitic and permits an omissible reciprocal complement (e.g. *unul pe altul* 'each other') such as in (6)a, the second accepts neither clitic nor reciprocal complement, (6)b.

(6) (a) Ion concura cu George în afaceri. ↔ Ion și George se concurau (unul pe altul) în afaceri.
John competed with George in business. ↔ John and George competed (with each other) in business.

(b) Viziunea regizorului concura cu cea a actorilor. ↔ Viziunea regizorului și a actorilor (*se) concurau (*una pe alta).
The director's vision concurred with that of the actors. ↔ The vision of the director and that of the actors concurred.

Unlike Kettnerová et al., we use neither general alternation rules, nor situational participants (e.g. Agent, Recipient, etc.). The distinction between the variants of the same alternation is done through semantic restrictions, because often an alternation does not imply a simple reorganization of the situational participants but a random complement variation, like in the example (7) for the verb *a achita* 'pay', where (a) and (b) are valency variants of the same meaning "to give money for a commercial product":

(7) (a) 1. NP [nom +person], 2. NP [acc +goods]
Ion a achitat băutura.
John paid the drink
'John paid for the drink.'
(b) 1. NP [nom +person], 2. NP [acc +money], 3. PP [pentru 'for' +goods]
Ion a achitat un euro pentru băutură.
John paid one euro for the drink.

Notice that in Romanian the commercial product can be expressed either by an NP in accusative (7) a when the paid price is not expressed, or by a *pentru* 'for'-PP (7)b, when the price is specified as direct object.

It is very likely that when DVCRL covers almost all the Romanian verbs one may obtain valuable generalizations about alternations by capturing the recurrent patterns in some rules.

3.4 Dependencies Between the Complements

In order to capture the dependencies between the complements of a verb pattern, expressing control or raising phenomena or case agreement, we have used a set of conventions.

If a verbal complement has, as its subject, the NP in nominative (that is, the same subject as the lemma verb), its part-of-speech is indicated simply by V, like in example (8) for the subject-control verb *a promite* 'promise':

(8) 1. NP [nom +person], 2. V [să 'to'], 3. NP [dative + person]
Ion promite Mariei să o iubească întotdeauna.
Ion.nom promises Maria.dat to her.clitic.acc love forever.
'Ion promises Maria to love her forever.'

The verb *promite* has the same subject as its verbal complement in subjunctive (marked with V[*să* 'to']): *să iubească* 'to love', namely the NP in nominative: *Ion*, while the NP in dative: *Maria* shows to whom is addressed the promise.

If the verbal complement has as its subject an NP different from the one in nominative, then we use VP (without any semantic restriction), like in example (9) for the object-control verb *a recomanda* 'recommend':

(9) 1. NP [nom +person], 2. NP [dat +person], 3. VP [să 'to']
Ion îi recomandă Mariei să plece.
Ion recommends to Maria to leave.

This time, the subject of the verb in subjunctive *să plece* 'to leave' is the person to whom the recommendation is addressed, namely the reference of the NP in dative.

When a complement is a subordinating clause that can have a subject that is different from the other NP-complements, we use the mark VP as well, but we also indicate a semantic restriction in the set +fact, +question, +cause, +goal etc. for that clause. The example (10) shows such a case.

(10) 1. NP [nom +person], 2. NP [acc +money], 3. (fac.) VP [+goal]
Ei cheltuie bani ca el să aibă succes.
They spend money for him to be successful.

The situation of case agreement concerns cases in which a complement is a predication of another complement, like in example (11) for the verb *a considera* 'consider, regard as', where the argument position 3 refers to the argument position 2 and must have the same grammatical case, here the accusative. These are structures of small clauses. The predication can be expressed by an adjective phrase (AdjP) or by a noun phrase, as variants of the same position.

(11) 1. NP [nom +person], 2. NP [acc], 3. AdjP [acc] / NP [acc]
Eu îl consider pe el bun/prieten.
I consider him good/my friend.

3.5 XML Encoding

Any XML (eXtensible Markup Language) encoding needs a DTD (Document Type Definition) that is a schema ensuring a logical and consistent structure of all the dictionary entries. The TshwaneLex DTD editor allows lexicographers to tailor the appropriate schema for every kind of dictionary, without the need for an IT expert. In Figure 3 we give a sketch of the DTD used in DVCRL and an XML example. The main lines in designing a DTD concern the information types captured in the encoding (identified by the element names), their embedding structure (see the indentation expressing a child relation) and the multiplication number of the corresponding elements (see the superscript symbols). An element encompasses other elements (with internal structure) and/or attributes (with terminal values).

In DVCRL schema, an entry, named Lemma, has ArgUnit and Expression as child elements and the attribute LemmaSign taking the verb lemma as its value. The ArgUnit element, which has to be at least one or more (see superscript +), encodes the verbal context and has a numbering attribute ArgUnitNr. The children of ArgUnit are the morphological information, which can appear once or can be missing (LemmaFeature with superscript (0,1)), one or more verb patterns (ArgStructure^{+}) and one or more numbered senses (Sense^{+}). Further, each element has the structure displayed in Figures 1 and 2. Note that the meanings of the superscript symbols are the following: * – zero or more, 1 – strictly one, and inside brackets there are lists of possible attribute values.

The TshwaneLex editor offers the possibility of assigning specific layout characteristics to DTD elements and attributes, thus relieving lexicographers of the need to worry about these. This facility can trigger, sometimes, the need for element restructuring and small schema modifications. For instance, for treating several expressions of a lemma as items of the same type we should include them into an Expression List element. In other words, the final form of a DTD can be motivated not only by the information structure, but by layout features as well. In the second column of Figure 3, one can see how DTD is reflected in XML encoding, by the example of the first verbal context of the verb *concura* (see verbal context I in Figure 1).[5]

5 The XML variant of this dictionary can be obtained by request from the author.

DTD schema	XML example for *concura* I.
Lemm:LemmaSign = *text* ArgUnit^{+}: ArgUnitNr = *integer* LemmaFeature*: ReflForm$^{(0,1)}$= [*se, îşi, o, îl*] NegForm$^{(0,1)}$=[nu] FlexForm$^{(0,1)}$= [*3sg*] ArgStructure^{+}: LemmaFeature*: ReflForm$^{(0,1)}$= [*se, îşi, o, îl*] NegForm$^{(0,1)}$=[nu] FlexForm$^{(0,1)}$= [*3sg*] ArgPosition^{+}: PositionNr=*integer* PositionState$^{(0,1)}$:PositionState=[fac., opt.] Argument^{+}: POS^{1}=[GN, GAdj, GAdv, GP, GV, V] Afeature^{1}: Gram*= *text* Sem*=*text* Sense^{+}: SenseNumber=*integer* Definition^{+}: Definition^{+}=text Example^{+}:Example^{+}=*text* Expression*: LemmaSign=*text* Definition=*text*	<Lemma LemmaSign="concura" <ArgUnit ArgUnitNr="1"> <ArgStructure > <ArgPosition PositionNr="1"> <Argument POS="GN"> <AFeature Gram="nom" Sem="+persoană / +animat"/> </Argument> </ArgPosition> <ArgPosition PositionNr="2"> <Argument POS="GP"> <AFeature Gram="la"Sem="+competiție"/> </Argument> </ArgPosition> </ArgStructure> <Sense SenseNumber="1"> <Definition Definition="a participa la un concurs"/> <Example Example="Şcoala noastră concurează şi ea (la campionatul liceelor)."/> </Sense> </ArgUnit> [...] </Lemma>

Figure 3: XM Encoding.

4 Challenging Aspects

The most challenging task in this approach was to establish the appropriate characteristics of a verb pattern, especially regarding the semantic features (or selectional restrictions) of each complement. This has been done by resorting to corpus examples but also to human intuition, because it is not always possible to reach, in a huge corpus like the internet, all the structures that are in use, especially those less frequent ones. Actually, a dictionary like DVCRL is comprised not of verb patterns found in corpus, but what it is *possible* to be found, described in a concise manner. We tailor an intuitive verb pattern, by starting from a meaning of a verb, and then verify it on corpus examples. After gathering the examples, the question that arises is what the complements have in common. Thus, certain semantic preferences can show up and they have to be defined in the most general way as possible. At the beginning we chose intuitive semantic tags such as +human, +object, +authority, +event, etc. We then extended this list as needed, although using the tags already defined as much as possible, in order not to increase the list excessively. At the end, we removed the semantic features assigned to less than three verbs and included them explicitly in the verb definition or as a note. For instance, one of the meanings of *a se ambala* is "get running uncontrollably fast" and requires a subject with the semantic feature +horse. Because this feature was singular, we have replaced it with +animate and specified in the definition that it is about a horse. Finally, we obtained a list of 90 semantic tags (for about 2,000 meanings).

Many verb patterns differ only by the semantic features of their complements. For instance, in Romanian the verb *amputa* 'amputate' has the medical use (of cutting off a limb) but also a general or

metaphorical meaning, namely "to remove a part of something with bad consequences" like in the example "[...] but people apparently have no problem with billions being squandered on an *amputated budget* in Europe" (http://www.europarl.europa.eu/). The latter meaning applies to budget, texts, personality, future, and so on, to the effect that no specific semantic preferences can be delimited. These two meanings are assigned the verb patterns in example 12, respectively:

(12) (a) NP [nom +animate], 2. NP [ac +limb]
(b) NP [nom], 2. NP [ac]

As one can notice, the verb pattern in (12)b includes, somehow, that in (12)a. However, which is better: to conflate or to keep them separately? Such questions arose throughout our work.

5 Conclusion

This study describes a verb valency lexicon for the Romanian language – here valency is understood in a broad sense, corpus and user-oriented. This dictionary is the first endeavor at this level of comprehensiveness and consistency for this language, and it is intended to be a valuable resource for both computational linguistics and teaching Romanian to foreign students. Besides the entry description, we tried to bring out some aspects of our experience, such as the necessity to detach the meanings of verb patterns from those in an usual monolingual dictionary. We consider this fact to be important because, as we know, many tasks of word sense disambiguation use such monolingual dictionaries as references, which do not reflect the formal aspects found in a corpus, as a valency dictionary does, and thus the computational performance is diminished. We also touch problems less discussed in the literature, such as a new type of optional complements, the argument position concept based on co-occurrence test and the dependencies between complements of the same verb pattern.

The dictionary presented here is just a starting point. There is a lot of room for improving the informativeness and the consistency of its entries. Information about semantic roles, diatheses and collocation preferences waits to be added, as does increasing the number of entries. We also hope to get valuable feedback from the dictionary's users as soon as possible, in order to use this as supplementary guidelines for further work.

References

Allerton, D. J. (1975). Deletion and proform reduction. In *Journal of Linguistics,* 11, pp. 213-238.

Barbu, A.M. (2008). First Steps in Building a Verb Valency Lexicon for Romanian. In P. Sojka et al. (eds.), *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligece (LNAI) 5246, Springer, pp. 29-36.

Barbu, A.M. (2017). Dicționar de contexte verbale. Editura Universității din București.

Barbu, A.M. & Ionescu, E. (2007). Designing a Valence Dictionary for Romanian. In *Proceedings of International Conference Recent Advances in Natural Language Processing* (RANLP2007), 27-29 September 2007, Borovets, pp.41-45.

DEX (1998). Dicționarul Explicativ al Limbii Române. Institutul de Lingvistică „Iorgu Iordan –Al. Rosetti" al Academiei Române, Editura Enciclopedic Gold.

Engel, U. & Schumacher H. (1976). Kleines Valenzlexikon deutscher Verben, Tübingen: Narr.

Fillmore, C. (1986). Pragmatically Controlled Zero Anaphora. In *Proceedings of the 12th Annual Meeting of the Berkeley Linguistics Society*, BLS12, Berkley, 15-17 February 1986, pp. 95-107.

Hanks, P., Pustejovsky J. (2005). A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, X-2, pp. 63-82.

Hanks, P. (2006). The Organization of the Lexicon: Semantic Types and Lexical Sets, http://www.cs.cas.cz/semweb/download/06-11-hanks.doc [03/20/2018].

Hartmann, Iren & Haspelmath, Martin & Taylor, Bradley (eds.) 2013. Valency Patterns Leipzig. Leipzig: Max Planck Institute for Evolutionary Anthropology, http://valpal.info [2018-03-08].

Helbig, G. & Schenkel W. (1968). Wörterbuch zur Valenz und Distribution deutscher Verben, Leipzig: Enzyklopädie.

Herbst Th. (1999). English valency structures - a first sketch. In *Erfurt Electronic Studies in English,* 6, http://web-doc.gwdg.de/edoc/ia/eese/rahmen22.html [03/07/2018].

Herbst, Th. & Heath, D. & Roe, Ian F. & Götz, D. 2004. A Valency Dictionary of English. A Corpus-based Analysis of the Complementation Patterns of English Verbs, Nouns, and Adjectives. Berlin – New York: Mouton de Gruyter.

Johnson, C. & Fillmore, C. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics* (ANLP-NAACL 2000), Seattle WA, pp. 56-62.

Kettnerová, V., Lopatková, M. & Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In *Proceedings of the 15th Euralex International Congress* 2012, 7-11 August 2012, University of Oslo, Norway, pp. 434-443.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. In *Computational Linguistics*, 29(3), pp. 333-347.

Kipper, K. & Dang, H. T. & Palmer, M. (2000). Class-based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence* (AAAII-2000), Austin, TX, 30 July-3 August, pp. 691-696.

Levin, B. (1993). English Verb Classes and Alternations. A Preliminary Investigation. Chicago and London: The University of Chicago Press.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F. & Świdziński M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavík, 26-31 May 2014, pp. 2785–2792.

Žabokrtský, Z. & Lopatkova, M. (2007). Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. In *The Prague Bulletin of Mathematical Linguistics,* 87, pp. 41-60.