# Aligning the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs

**Ana-Maria Barbu[1,2], Verginica Barbu Mititelu[3], Cătălin Mititelu[1]**

[1]"Iorgu Iordan – Al. Rosetti" Institute of Linguistics, [2]University of Bucharest,
[3]Romanian Academy Research Institute for Artificial Intelligence
13 Calea 13 Septembrie, Bucharest, Romania
anamaria.barbu@g.unibuc.ro, vergi@racai.ro, catalin.mititelu@lingv.ro

**Abstract**

We present here the efforts of aligning two language resources for Romanian: the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs: for each occurrence of those verbs in the treebank that were included as entries in the lexicon, a set of valence frames is automatically assigned, then manually validated by two linguists and, when necessary, corrected. Validating a valence frame also means semantically disambiguating the verb in the respective context. The validation is done by two linguists, on complementary datasets. However, a subset of verbs were validated by both annotators and Cohen's $\kappa$ is 0.87 for this subset. The alignment we have made also serves as a method of enhancing the quality of the two resources, as in the process we identify morpho-syntactic annotation mistakes, incomplete valence frames or missing ones. Information from each resource complements the information from the other, thus their value increases. The treebank and the lexicon are freely available, while the links discovered between them are also made available on GitHub.

**Keywords:** treebank, verbal valence frames, alignment, Romanian, semantic disambiguation

## 1. Introduction

Among the language resources for Romanian, there are ones that describe the syntactic and semantic aspects of the language. Two of them are the Romanian Reference Treebank[1] (henceforth RRT) (Barbu Mititelu, 2018) and the Valence Lexicon of Romanian verbs[2] (henceforth DCV) (Barbu, 2017). The former is a source of possible syntactic structures in the language, complemented by morphological and lexical information, as well as genre specifications, but no semantics can be found here. The latter resource is an inventory of verbal subcategorization frames designed for different verbal senses (clustered together when possible), complemented with as much as necessary morphologic information. We present here the work for coupling these two resources, thus increasing their value: information from each resource complements the information from the other: a verb in DCV, where its semantics and subcategorization frames with examples are presented, is now presented "in action": its occurrences in RRT show how the frames are lexicalized, how contextual elements can sometimes prevent some valences from being lexicalized. Given that DCV was created starting from examples in a journalistic corpus, they are now tested against occurrences in other genres, those represented in RRT (see below). Vice versa, verbs in RRT are associated with possible senses as defined in DCV, are grouped together under the same frame irrespective of the variations of their syntactic structures favoured by the context of occurrence.

After presenting related work on aligning a treebank and a valence lexicon in Section 2, we describe the two Romanian resources (in Section 3) and the methodology adopted for aligning them (in Section 4). The results obtained are presented and discussed in Section 5. The way in which this alignment contributed to enhancing the quality of the two resources is described in Section 6. A discussion of problematic cases that challenge the alignment process can be found in Section 7, just before concluding the paper.

## 2. Related Work

A project that aims, like our project, to correlate a corpus with a valence lexicon and to correct the respective resources is presented by Woliński and Hajnicz (2021). They aim to harmonize the Polish treebank Składnica with the Walenty dictionary through the Swigra 2 parser, bringing improvements to all 3 resources. The research focuses on the syntactic side and aims to unify the syntactic functions and take over as much information from Walenty as possible in Składnica. Thus, adverbial-like arguments were detailed from old advp in 10 specific subtypes of xp (expressing time, duration, place, starting or ending point, etc.); special types of arguments present in Walenty have been implemented, e.g. complex prepositions; solutions have been found for the representation of "unlike coordination" (coordination between arguments from the same position that have different syntactic types), the representation in the treebank of the discontinuous constituents and so on. Unlike this project, which focuses on the syntactic part, our project achieves for the analyzed verbs especially a word sense disambiguation, because when a verb in RRT is assigned a certain valence frame in DCV, it is also assigned the corresponding meaning.

Another example of a treebank coupled with a valence dictionary is presented by Hinrichs and Telljohann (2009). Here, unlike our project, a correspondence is made between the dictionary complements of a verb and those in the treebank TüBa-D/Z, using the same morpho-syntactic labels, but the semantic information is missing.

PropBank (Palmer et al., 2002), on the other hand, is an annotated morpho-syntactic corpus to which annotation with argument structures expressed by semantic roles has been done manually. From PropBank, Cinková (2006) extracted the EngVallex Dictionary of Valences, with which the corpus is coupled. EngVallex, according to the model of the Czech valence dictionary Vallex (Lopatková,

[1] https://github.com/UniversalDependencies/UD_Romanian-RRT
[2] http://188.212.37.221:9000/

2003), contains both morpho-syntactic and semantic information, and can be linked with other corpora (see Cinková, 2006).

On the line of disambiguation of valence frames, there are experiments around VALLEX and VALEVAL, as presented by Lopatková et al. (2005). We refer to this work especially in terms of disambiguating the frames by human annotators, as in our project, in order to obtain a gold standard. But we have not yet developed an automatic tool trained for this purpose. An extensive alignment between Vallex and Prague Dependency Treebank was made by Hajič et al. (2003) and Urešová et al. (2014).

## 3. Resources Description

In this section we describe the two language resources for Romanian: RRT and DCV. We present their design principles and give some statistics of their content.

### 3.1. The Romanian Reference Treebank

A treebank is a syntactically annotated corpus (Abeillé, 2003). Besides the syntactic level, two other analysis levels are also present: lexical (sentences are tokenized and lemmatised) and morphologic (each word is morphologically disambiguated and the information is encoded in a part-of-speech (PoS) tag).

RRT was created following the Universal Dependencies[3] (UD) (Nivre et al., 2016; de Marneffe et al., 2021) guidelines for syntactic annotation[4]. Briefly, what is specific to this project is the fact that the morpho-syntactic annotation is meant to be cross-lingually valid and, simultaneously, to serve as training and testing material for the development of multilingual parsers. The annotation principles are established so as to serve these objectives: "The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary."[5]. As far as syntax is concerned, each sentence is a tree, with the first main verb (in linear order) as its root. An exception to this is the case when the first main clause in a sentence contains the copula verb *be*: as explained below, the copula is a function word and, thus, its predicative is analysed as the root of the tree. Content words are given primacy, assuming that they, rather than function words, have equivalents cross-lingually. Function words (i.e., prepositions, auxiliary verbs, copula verbs, determiners, conjunctions) attach as dependents of content words, not of other function words[6]. Figure 1 shows the tree structure of the sentence in example (1) taken from RRT:

(1) Înmormântarea sa va avea loc la Sandhurst săptămâna
Funeral-the his will have place at Sandhurst week-the

viitoare.
next.

'His funeral will take place in Sandhurst next week.'

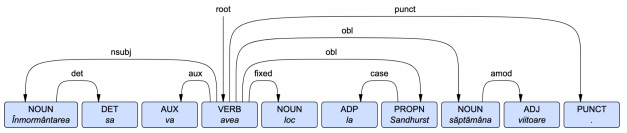


Figure 1: The tree structure of a sentence in UD format.

Figure 1 shows that the root of the tree is the main verb (and the only one here), its subject (*Înmormântarea*) is a nominal one (see the use of the relation `nsubj`), the verb and the following noun (*loc*) make up a fixed construction (a light verb construction in this case) (see the relation `fixed`), the verb has two oblique dependents (see the two `obl` relations), the end of sentence punctuation attaches to the root, while all function words attach to the content words: the possessive adjective *sa* is a determiner (`det`) of *Înmormântarea*, the auxiliary verb *va* attaches to the main verb (see the relation `aux`), the preposition *la* enters the `case` relation with the oblique *Sandhurst*, while the other oblique is modified by the adjective *viitoare* (see the relation `amod`).

As shown in Figure 2, in UD there are core arguments and non-core dependents of a verb, both types being further distinguished with respect to their lexical or clausal lexicalization: `nsubj` (nominal subject), `csubj` (clausal subject), `obj` (usually the direct object, i.e. "the entity acted upon or which undergoes a change of state or motion"[7]), `iobj` (usually the indirect object, i.e., the recipient), `ccomp` (the clausal core argument corresponding to the lexical `obj` or `iobj`), `xcomp` (open clausal complement, i.e., a predicative or clausal complement without its own subject). None-core dependents are: `obl` (nominals that are adverbials of the verb), `advmod` (adverbs that are adverbials) and `advcl` (as the clausal correspondent of the other non-core dependents).

[3] https://universaldependencies.org

[4] RRT is one of the four Romanian treebanks available in UD: the others are Romanian Nonstandard (Mărănduc and Bobicev, 2017) (a treebank reflecting the nonstandard language), SiMoNERo (Barbu Mititelu and Mitrofan, 2020) (a medical treebank), and ArT (Barbu Mititelu et al., 2021) (a treebank of the Aromanian dialect of the Romanian language). We have chosen RRT for our endeavour because it reflects the standard use of the language and covers several genres.

[5] https://universaldependencies.org/introduction.html

[6] The exception to this is when function words are coordinated with each other, as in "The freshness and unique taste of our chocolates remain intact *until and after* the chocolate delivery" (https://www.ovidias.com/en-ie/payment, accessed 6th Jan 2022), where the preposition *until* is the dependent of *delivery*, while the conjunction *and* and the preposition *after* are dependents of *until*.

[7] https://universaldependencies.org/u/dep/obj.html

| | Nominals | Clauses | Modifier words | Function Words |
|---|---|---|---|---|
| Core arguments | nsubj<br>obj<br>iobj | csubj<br>ccomp<br>xcomp | | |
| Non-core dependents | obl<br>vocative<br>expl<br>dislocated | advcl | advmod*<br>discourse | aux<br>cop<br>mark |

Figure 2: The UD relations whose head can be a verb (source: https://universaldependencies.org/u/dep/index.html).

With respect to `obj` and `iobj` relations, it is worth mentioning that, even though they are rather clearly distinguished from each other semantically (i.e., `obj` is the patient and `iobj` is the recipient) and morphologically (`obj` is in the accusative case, while `iobj` is in the dative case, for languages with case, like Romanian), in the situation of ditransitive verbs with two accusative objects (e.g., *teach somebody something*), the UD convention is to annotate the [+Animate] one as the `iobj` and the other as the `obj`[8]. The same annotation convention applies in examples with a raised argument: see example (7) at point f) in section 7 below: the verb *împiedică* combines with a subject and a direct object: when the direct object is syntactically realized as a subordinate clause (annotated with the relation `ccomp` in UD), its subject is raised in the main clause: given that the relation `ccomp` is the clausal counterpart of the relation `obj`, they cannot co-occur with the same verb, so the raised object should be annotated as `iobj`, in spite of its morphologic and semantic characteristics which are not those specific for `iobj` (described above).

Coordination is treated asymmetrically in UD, with the first conjunct (in linear order) as the head of the other conjuncts, while the (coordinating) conjunctions are dependents of the conjunct they precede. Whenever conjuncts share dependents, the latter are attached to only one of the conjuncts, as shown in Figure 3, where we show the tree structure of the sentence in example (2): the verbs *reproduc* and *modifică* share the same direct object (*întâmplările*), but it attaches only to the first (in linear order) of the conjuncts and there is currently no mechanism in UD of retrieving the information that it is actually shared by both.

(2) ei reproduc sau modifică întâmplările auzite sau citite.

they reproduce or modify happenings-the heard or read

'They reproduce or modify the heard of or read happenings.'

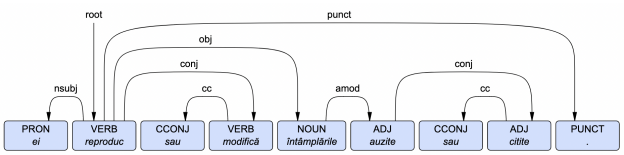


Figure 3: The tree structure of a sentence containing coordination.

[8] https://universaldependencies.org/ro/dep/iobj.html

UD allows for postulating language-specific relations and, for Romanian, we mention here `obl:pmod` (for the prepositional dependents whose preposition is selected by the verb, such as the preposition *on* in the sentence "Everything depends on this.", but not as in the sentence "I sat on the chair."); `ccomp:pmod` (for clausal lexicalizations of such prepositional dependents); `obl:tmod` (for temporal nominal modifiers); `adv:tmod` (for temporal adverb modifiers); `advcl:tmod` (for temporal clausal modifiers).

A relation specific to various uses of clitics is `expl`. In Romanian it is used for the clitics doubling the direct object, the indirect one, as well as the subject, but also for the non-referential use of pronouns (such as the pronoun *o* in *a luat-o la dreapta* has taken-it at right 'he has turned to the right'). There are also four relations subtypes of `expl` that are used for different values of the reflexive clitic: `expl:pv` (for its inherently reflexive value), `expl:impers` (for its impersonal value), `expl:poss` (for its possessive value) and `expl:pass` (for the passive value).

Many verbal idioms have a flat analysis in RRT, being annotated with the relation `fixed` (see example (1) and Figure 1), as they may raise difficulties when syntactically parsed.

RRT contains 9,523 sentences, 218,511 tokens and 17,278 unique lemmas. The average sentence length is 23 tokens. The sentences belong to different genres, clearly marked in the files: fiction − 1,818 sentences, law − 1,606 sentences, medical − 1,210 sentences, FrameNet translations − 1,092 sentences, academic writing − 950 sentences, news − 933 sentences, science − 362 sentences, wikipedia − 251 sentences, miscellanea − 1,301 sentences. The treebank is released within UD.

### 3.2. The Valence Lexicon

The valence lexicon (DCV, version 1.15) contains 628 Romanian verbs for which 2,372 valence frames were created and 2,476 (sub-)senses defined, manually. Of these verbs, 486 (i.e. 77%) are polysemous, that is, they have at least 2 valence frames and the average degree of polysemy is 3.78 frames per verb. Multiple (sub-)senses can be assigned to the same valence frame, just as multiple valence frames (alternating) can have the same meaning(s). The valence frame contains both the obligatory complements and the optional complements that are closely related to the meaning of a verb and are frequently used with that verb, somewhat similar to the frame-semantics approach in FrameNet (Fillmore and Atkins, 1992) and PropBank. To suggest this extension of the frame to its minimal form, the lexicon was called Dicționar de Contexte Verbale 'Dictionary of Verbal Contexts' (shortly DCV).

Example (3) illustrates the valence frames of the verb *adăuga* 'add':

(3)

| **Frame #** | **I** |
|---|---|
| Complements | 1. GN [ nom ] 2. GN [ ac ] 3. (fac.) GP [ +loc ] / GN [ dat ] |

| | |
|---|---|
| Meanings | a pune în plus<br>'to put in addition' |
| Examples | Școala suceveană și-a adăugat încă patru medalii în palmares.<br>'The Suceava school added four more medals to its record.' |
| **Frame #** | **II** |
| Complements | 1. GN [ nom +persoană ] 2. GN [ ac +text ] / GV [ că +text ] / GV [ - +text ] |
| Meanings | a spune sau a scrie ceva în completare<br>'to say or write something in addition' |
| Examples | Alex a adăugat o știre nouă.<br>'Alex added a new story.'<br>Premierul a adăugat că guvernul caută totuși soluții.<br>'The prime minister added that the government was still looking for solutions.' |

The description of the valence frames of a verb starts from the meanings assigned to the respective verb in the explanatory dictionary of the Romanian language, making inevitable adaptations. Because DCV was designed not only for human use but also as an NLP resource, the description focuses largely on formal marks detectable in text relatively easily (through primary processing), such as parts of speech (e.g. as heads of GN = noun phrase, GP = preposition phrase, GV = verb phrase / sentence), grammatical cases (e.g. nom = nominative, ac = accusative, dat = dative), lexical marks: conjunctions (e.g. *că* 'that'), some prepositions, etc. Differentiation of valence frames can sometimes be done only by semantic restrictions (or preferences) that are specified in some complements and are marked with '+' (e.g. +persoană 'person', +text 'text', +loc 'location'). Semantic restrictions can be defined with senses from a wordnet (Miller, 1995; Fellbaum, 1998), but in this project they have been ignored. As can be seen in example (3), no syntactic functions are used, nor distinctions between complements and adjuncts (however, optional complements are marked with (fac.)) or semantic roles. We have abandoned such controversial information in order to simplify the construction of the DCV and to make it as flexible a resource as possible.

In addition to the valence frames, a verb can be associated with a list of glossed expressions, headed by the verb. Expressions can be considered fixed valence frames, with lexicalized complements, in which the verb loses its meaning in favor of the meaning of the whole expression. A more detailed description of DCV is given by Barbu (2018), and its XML format and related DTD are available at http://188.212.37.221:9000/.

## 4. Methodology

The task of aligning the two resources can be described as: for each occurrence of those verbs in RRT that were included as entries in DCV, the valence frame is automatically searched for among those recorded for the respective verb in the lexicon, then manually validated and, when necessary, corrected. Several steps were taken to this aim and they are described in the subsections below.

### 4.1. Step 1: Define the pool of verbs

The verbs of interest for us are, first of all, those main[9] verbs occurring in both resources. Secondly, after evaluating the syntactic behaviour of some lemmas, we decided to exclude them: this is the case of the verbs *putea* 'can', *trebui* 'must' and *fi* 'be'. The first two were excluded because they act like modal verbs, but they were not annotated as such in RRT, while the verb *fi* can be either a function word (an auxiliary or copula) or a main verb, in which case it is sometimes very difficult to distinguish among its meanings. All main verbs occurring in passive constructions (be they regular or reflexive passives) were left out, because this construction presupposes a reorganization of the subcategorization frame and this was beyond our interests in this endeavour. All passive constructions could be easily identified in RRT: the regular passive is characterized by the presence of an auxiliary annotated with the specific relation `aux:pass`; the reflexive passive is characterized by the presence of a reflexive clitic which is annotated with the relation `expl:pass`[10]. A reorganization of the verbal valences also occurs with participles and supines, thus verbs occurring in these moods were also left out. The number of verb occurrences considered for our alignment was 12,198 (accounting for 567 verbs from DCV).

### 4.2. Step 2: Define the mapping table between the relations in RRT and the valences in DCV

A list of mappings between the UD relations of interest and the complements defined in the valence frames was created. In Table 1 there are only some of the mappings we have defined.

| UD relation | DCV complement |
|---|---|
| nsubj | GN [nom] |
| csubj | GV |
| | V [să] |
| obj | GN [ac] |
| iobj | GN [dat] |
| ccomp | GV |
| obl<br>obl:tmod | GP |
| | GAdv |
| | GN [ac] |
| obl:pmod | GP |
| expl:pv | V [se] |

Table 1: Mappings between UD relations and DCV complements. GN stands for noun phrase, GV for verb phrase, GAdv for adverb phrase and GP for prepositional phrase. The brackets contain case or form restrictions.

[9] Auxiliary verbs were, clearly, excluded.

[10] There are also subtypes of the relations marking the subject that occur only in passive constructions, namely `nsubj:pass` and `csubj:pass`, but, given that Romanian is a pro-drop language, they may not be present in the passive structure.

These mappings were used for writing the rules on the basis of which the algorithm tries to assign a valence frame from DCV to a verb's occurrence in RRT.

For the occurrences of verbs in RRT, the algorithm could propose one or more frames from DCV, it was not able to propose any frame or it recognized the use of that verb within an expression (based on the `fixed` relation the verb heads in RRT). Their distribution is presented in Table 2.

| # verb occurrences in RRT | 12,198 | |
|---|---|---|
| # verbs for which one or more frames from DCV was/were proposed | 9,992 | 82% |
| # verb occurrences for which no frame could be automatically proposed | 1,708 | 14% |
| # verb occurrences as parts of an expression | 498 | 4% |

Table 2: Possible outputs of the alignment algorithm.

An example of the mapping result is the following, where we boldfaced the validated mapping:

```
# sent_id = train-7768
# text = Intensitatea liniilor spectrale scade treptat pe măsură ce se micșorează lungimea de undă.
scade[4] (scădea):
   unit: 1
     Intensitatea: nsubj -> GN [nom]
     treptat: advmod -> GP [] / GAdv []
     _: advcl:tcl -> _
   eval: c

   unit: 4
     Intensitatea: nsubj -> GN [nom]
     treptat: advmod -> GAdv []
     _: advcl:tcl -> _
   eval:
   DCV: n
```

### 4.3. Step 3: Validation of the alignments between verbs occurrences and valence frames

The validation phase was meant to cover all three alignment types presented in Table 2:

- proposed frames were either accepted or rejected;
- when no frame was proposed, we: (i) manually tried to propose one, (ii) assigned the occurrence to an expression, or (iii) confirmed that no frame was assignable, given its absence from DCV;
- expressions were only confirmed.

Manual validation, until the moment of this writing, has covered 7,192 verbal occurrences, i.e. 59% of the data, and it still continues. They are distributed as shown in Table 3.

| # verb occurrences in RRT | 7,192 | |
|---|---|---|
| # verbs for which one or more frames from DCV was/were proposed | 5,881 | 81.8% |
| # verb occurrences for which no frame could be automatically proposed | 987 | 13.7% |
| # verb occurrences as parts of an expression | 324 | 4.5% |

Table 3: Distribution of the alignment outputs in the data validated so far

The validation of the frames was done by two annotators on complementary parts. In order to verify the degree of reliability of the validation, we proceeded to the double validation of a sample of the data, obtaining the results in Table 4. These data show that for 219 verb occurrences out of 250 the two annotators assigned the same DCV frame independently and without any predefined protocol and the calculated Cohen's $\kappa$ is 0.87, showing a high agreement between them.

| # double-validated sentences | 109 |
|---|---|
| # double-validated verb occurrences | 250 |
| # number of verb occurrences with annotators agreement | 219 |
| Cohen's $\kappa$ | 0.87 |

Table 4: Results of the double validation

It is worth mentioning that the double validation covers 123 verb lemmas with an average distribution of 2.03 per sentence. A number of 19 lemmas attracted the disagreement of the annotators. The disagreement is not surprising in the case of light verbs, such as *face* 'do', *lua* 'take', *da* 'give', but the others, most of them (e.g. *auzi* 'hear', *spune* 'say', *rezulta* 'result'), are probably not very clearly defined in DCV.

## 5. Results

We will discuss the results of the validation phase for each type of the alignment outputs. Note that in this section by *verbs* we mean *verb occurrences* (not lemmas).

### 5.1. Validation of proposed frames

The alignment algorithm proposed (ambiguous or unambiguous) frames for 5881 verbs. For 5359 of these verbs (91%), a correct assignment could be manually identified. For the remaining 9% of the verbs, none of the proposed frames were correct.

In Table 5 we show that for about a tenth of the data (596 verbs) DCV has only one frame which was obviously selected by the algorithm. The fact that only 592 of them were manually validated is suggestive of the need to include more frames for the respective verbs in DCV.

Other unique frames were proposed for 2,325 verbs that have more frames defined in DCV and 2,100 (90%) of them were considered correct. These are usually cases when, for the same verb, the frames are syntactically or lexico-syntactically well distinguished: e.g., in one frame the second complement occurs in accusative, in another in dative or with a certain preposition. Whenever the syntactic characteristics of a complement does not differ from one frame to another and the rationale behind having distinct frames is only of semantic relevance, the algorithm cannot but propose all syntactically similar frames and it is the role of the human validator to distinguish between them. Any semantic disambiguation of the sentences in RRT could have probably helped the algorithm to find the right frame, but such information was not available.

| | | Validated frames | |
|---|---|---|---|
| **# verbs with proposed frames** | **5,881** | **5,359** | **91%** |
| **# verbs for which a unique frame was proposed** | **2,921** | **2,692** | **92%** |
| # verbs for which the unique frame proposed is unique in DCV | 596 | 592 | 99% |
| # verbs for which the unique frame proposed is selected out of two or more in DCV | 2,325 | 2,100 | 90% |
| **# verbs for which more frames were proposed** | **2,960** | **2,667** | **90%** |

Table 5: Results of the manual validation of the automatically proposed frames

### 5.2. Verbs with no proposed frames

In Table 6 we can see that for 681 verbs (almost 70% of the cases) for which no frame could be automatically assigned, the human validator was able to find the correct one in DCV. The automatic possibility of finding one was hindered by various factors, most of them being annotation errors in RRT: any incorrect syntactic relation precludes the mapping between the two resources.

For 45 cases (about 5% of the data), a frame could not have been found because, as the human validators showed, the verb was actually part of an expression which, given the known idiosyncrasy expressions manifest (Baldwin and Kim, 2010), did not match any of the defined frames, even if they had been incorrect ones.

There are still 261 cases (about 25% of the data) for which not even the manual validators could propose a frame from DCV and several causes were identified for this:

- incorrect part of speech (PoS) tags in RRT: e.g., the foreign noun *Fabrice* is identical in form (except for the capitalization of the first letter) with one form in the paradigm of the Romanian verb *fabrica* (En. "fabricate") and it was tagged as a verb instead of a noun;
- passive constructions incorrectly annotated as active ones in RRT;
- metaphorical use of verbs for which DCV does not contain defined frames;
- cases of homonymy: homonyms are treated as separate verbs and the algorithm tries to match the occurrence of a verb with any frame of either homonym in DCV, one of them being actually inapplicable;
- missing frames from DCV entries.

| **# verbs with no proposed frames** | **987** | |
|---|---|---|
| # verbs for which a frame was proposed manually | 681 | 69% |
| # verbs as parts of expressions | 45 | 5% |
| # verbs for which no frame can be proposed manually | 261 | 26% |

Table 6: Solutions proposed in the manual validation step for the verbs lacking an automatically assigned frame

### 5.3. Verbs within expressions

A number of 324 verb occurrences were recognized by the algorithm as part of an expression because they are heads of at least one `fixed` relation in RRT. To this number we need to add another 45 occurrences that were identified manually while going through the set of verb occurrences for which the algorithm could not propose any frame from DCV: these cases show the inconsistent annotation of verbal expressions in RRT and they required correction.

## 6. Resources Corrections

Following manual validation, corrections were required in both resources and this was one of the important aims when proceeding to this endeavour.

### 6.1. RRT

RRT was developed with a lot of manual work, indeed helped by an iterative bootstrapping mechanism: after manually syntactically annotating 500 sentences, they were used as training material by a parser, which was subsequently run on another set of 500 sentences. The resulting annotation was manually corrected and these sentences enriched the original set of sentences used as training material, and the parser was run again on the enriched set. This cycle repeated until all 9523 sentences were corrected. However, no consistency check has ever been run on RRT, thus, the activity reported here serves as a method of spotting existing annotation errors and correcting them. Given that simultaneously RRT underwent corrections as a result of its use in the development and tuning of a neural networks part of speech tagger for Romanian, we can report here on the types of errors we have corrected during the validation step, but not on the number of each type of corrections.

Many errors that have been corrected concern the passive structures:

- they had been marked inconsistently: e.g.: the subject of the verb is assigned the `nsubj:pass` relation, which is specific to passive subjects, but the auxiliary is not assigned the `aux:pass` relation (specific to such constructions), but `aux`;
- they were not marked at all, which means they were probably not recognized as such;
- they were wrongly identified as passives instead of impersonal, for instance, when they are lexicalized in the form of a reflexive passive construction.

There were also cases of mistaking the `nsubj` for `obj` or vice versa. This could have happened because of the homonymy of the nominative (the default one for subjects in Romanian) and accusative (the one specific for direct objects) cases in Romanian, because of the relatively free word order of Romanian and, rarely, of the ambiguous structures.

Reflexive pronouns are sometimes very difficult to assign a certain semantic value: passive and impersonal values may be impossible to clearly tell apart in some contexts, given that the context represented by the sentence is not enough for disambiguating them.

As already mentioned above, verbal expressions that were not annotated but were identified during the manual validation have also been annotated using the `fixed` relation in RRT.

### 6.2. DCV

Because DCV was built manually, the approach presented in this paper is the first validation by confronting the lexicon with a corpus.

In DCV we operated more than 400 modifications in all aspects. Here are some examples:

- We have introduced 17 new verbs, which are not among the most frequent in the language, but are quite frequent in RRT e.g. *prezenta* 'present', *conține* 'contain', *răni* 'wound', *întinde* 'extend', *întoarce* 'return', etc.
- We have added new valence frames corresponding to specific meanings found in the text. For example, the verb *ascunde* 'hide', when having a concrete meaning, admits a complement introduced by any locative preposition (e.g. *John hid the ball under/on/behind the table*), while the abstract meaning admits, in Romanian, only the preposition *sub* 'under': *Al. Roman s-a ascuns* sub *pseudonimul Cassius* 'Al. Roman hid under the pseudonym Cassius'; *Politicienii își ascund propriile eșecuri* sub *o retorică mincinoasă* 'Politicians hide their own failures under false rhetoric'. This determined us to separate the concrete meaning from the abstract one, each with its own valence frame, the abstract meaning having a strict prepositional regime (*sub* 'under').
- Numerous expressions and subsenses have been added, matching those found in the corpus.
- Some semantic restrictions have been removed if they have proven too restrictive, and others have been added. Although the automatic alignment did not address semantic aspects, manual validation took them into account, as semantic restrictions may help to select the appropriate valence frame in the context of several syntactically identical frames. For instance, the verb *absorbi* 'absorb' has two identical syntactic valences: a subject and a direct object, but the semantic restrictions differentiate two very different meanings. The general meaning of *absorbi* is "incorporate" (*The sponge absorbs water.*), but if the direct object is +human, then the verb is a psychological one with the meaning "engross" (*The study of da Vinci's painting absorbed John.*).

## 7. Disagreements between RRT and DCV

As presented above, the alignment tool could not assign the correct valence or any valence to some of the verbs in RRT. These cases fall largely into one of the following types, due either to the disagreement between the linguists or to the differences between the convention systems adopted for the two resources, RRT and DCV.

### 7.1. Disagreement between linguists

a) **The different interpretation of the reflexive clitic *se*.** This pronoun can be: argumental (annotated as `obj` in RRT: *Ion* se *spală*. 'John washes himself') or non-argumental (without a corresponding valence in the frames in DCV, but marked as a morphological characteristics of the verb form when it is either inherently or contextually reflexive) in the following situations:

- it is obligatorily inherent for the so-called reflexive verbs: *Ion* se *uită pe geam* 'John SE looks out the window' – in RRT the clitic is annotated with the relation `expl:pv` in such examples;
- it is contextually inherent for transitive verbs that can also have meanings of change of state: tranz. *Soarele arde vopseaua* 'The sun burns the paint' / refl. *Vopseaua* se *arde de la soare* 'The paint SE burns from the sun.' – in RRT the clitic is annotated with the relation `expl:pv` in such examples;
- it is a marker of passive constructions: *Cărțile* se *vând în librării.* 'Books SE /are sold in bookstores.' – in RRT the clitic is annotated with the relation `expl:pass` in such examples;
- it is a marker of impersonal constructions: *Se merge prea repede.* SE go too fast / 'It's going too fast'. – in RRT the clitic is annotated with the relation `expl:impers` in such examples.

The only case that does not pose problems is the one when the clitic marks inherently reflexive verbs, but the others can be confused with each other, yielding disagreement among linguists.

b) **The dative of the Beneficiary.** Complements in the dative are in most cases argumental (in RRT: `iobj`). However, when they have the role of Beneficiary, they can appear unpredictable depending on verbs that do not include it in their valence. This is the case in example (4) where the reflexive dative *își* appears as a complement of the verb *găsi* 'find':

(4) Winston își găsise un loc pe jos.

'Winston himself.dat had found a place on the floor.'

It is worth mentioning that in Romanian the beneficiary dative competes with the possessive dative (which is non-argumental) and the two can be confused.

### 7.2. Representation differences in RRT and DCV

a) **Suppression, in context, of mandatory complements.** In DCV the obligatory complements are those that cognitively cannot be suppressed. For example, even if one can say "John is eating now.", everybody knows that John is actually eating something, so that the verb *eat* is always transitive cognitively. In real texts, some mandatory complements are not expressed, being implied in the context, which makes the alignment tool miss the selection of the appropriate frame. Example (5) lacks the place complement of the verb *intra* 'enter':

(5) În sfârșit, se deschise ușa și bătrânul intră.

'Finally, the door opened and the old man entered.'

b) **Lack of complements of the second verb in a coordination in which the verbs share their complements**. By convention, in RRT if two verbs share the same complement, this is only a dependent of the first of the verb conjuncts, and the second one wrongly seems to have no complements. In example (2) above, the alignment tool misses the transitive valence of the verb

*modifică* because in RRT the noun *întâmplările* is not in `obj` relation to it, but only to the verb *reproduc*.

c) **Cross-dependencies.** In RRT the dislocated complements are not related to the verb on which they depend semantically, but to the head of this verb, in order to avoid cross-dependencies. In example (6) *căreia* 'to which' is actually the argument of the verb *corespunde* 'fit', but in RRT it is annotated as `iobj` of *făcea* 'made'.

(6) categorie căreia Edgar Poe îi făcea să corespundă câmpul strict al poeziei

'category to which Edgar Poe made the strict field of poetry fit'

This type of problem also includes the phenomenon called clitic climbing, which exists in Romanian and which involves raising the clitic pronoun of a verb on its modal or auxiliary governor. For example, in the structure *Colesterolul se poate acumula în pereții arterelor.* 'Cholesterol can build up in the walls of the arteries.' the reflexive pronoun *se* depends semantically on the verb *acumula* 'build up', but syntactically, as in RRT, it depends on *putea* 'can'.

d) **The direct object of the raising verbs.** The raising verbs have, within the valence frame, a direct object and a clausal complement whose subject, not expressed in the subordinating clause, is in fact the direct object of the raising verb. In example (7):

(7) Frigul împiedică copacii să înflorească.

Cold-the prevents trees-the.acc.pl SĂ flower.vb.subj.3pl.

'The cold prevents the trees from flowering.'

the raising verb *împiedica* 'prevent' has the direct object (in accusative) *copacii* 'the trees' and the clausal complement with the verb in the subjunctive mood *să înflorească* 'that flower', whose subject is *copacii*. In RRT, according to an UD convention, in such situations, the direct object is not marked with `obj`, but with `iobj` (specific to complements in dative), whereas the clause has the `ccomp` label. This obj→iobj change prevents the alignment tool from assigning the correct valences to raising verbs.

e) **The direct object of the *dicendi* verbs.** In RRT, there are a lot of expressions of this kind: "How beautiful you are, he said!" or "I don't know, she answered". The texts "how beautiful you are" and "I don't know" are what is actually uttered and they are semantically the clausal core-arguments of the dicendi verbs *say* and *answer*, respectively. Due to the paratactic relation, these clauses are not marked as complements of the dicendi verbs, in RRT, and so the appropriate valence frame can not be recognized for them.

f) **Particularities of the text.** For some wording in the text, there is no valence frame/meaning in the DCV, because 1) they have a metaphorical use, for instance *a se deschide* 'open' in association with *izvor* 'water spring': *Unde dormeam se deschideau izvoare.* 'Where I slept springs opened.'; or 2) they represent mispronunciations, e.g. *spera* 'hope' requires the preposition *la*, which in a sentence from RRT is missing: *Se poate spera *(la) eliminarea spontană a pietrei la rinichi.* 'Spontaneous removal of kidney stones can be expected.'.

## 8. Conclusions and further work

This paper presents a new stage in the development of two important resources for Romanian language: RRT, the first dependency treebank of a considerable size and diversity, and DCV, the first valence lexicon with very high degree of detail. The corrections made in these resources will have beneficial implications for machine learning applications. RRT and DCV are valuable training and testing material for the development of dependency and disambiguation parsers and tools. In this direction, sentences with aligned verbs are a corpus that can be used as a gold standard for automatic disambiguation of valence frames. On the other hand, the alignment of the two resources makes it possible to enrich each with information from the other.

Our work has highlighted issues that make alignment difficult. Some can be improved (for example, by setting criteria by which linguists resolve their disagreement), but others cannot be improved (for example, system conventions).

It is encouraging that the alignment tool (which can be further improved) has managed to provide a unique valid frame for 50% of occurrences (see Table 5), only on the basis of syntactic information and despite the high degree of polysemy of verbs in DCV. From a lexicographical point of view, this result emphasizes the importance of valences in distinguishing senses.

The next steps in our work are to complete the validation for the entire RRT and try to align the complements within the valence frames. This latter step will open the way for annotating DCV with syntactic functions and then annotating both resources with semantic roles.

Both resources will be converted to Linked Data format (Chiarcos et al., 2013), which offers solutions for storing links between different resources. However, for the moment, the alignments we have validated can be downloaded from GitHub[11].

## 9. Acknowledgements

Part of the work reported here has been carried out within Action 2020-EU-IA-0088 which has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278547.

## 10. Bibliographical References

Abeillé, A. (Ed.) (2003). *Treebanks. Building and Using Parsed Corpora*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Baldwin, T. and Kim, S.N. (2010). Multiword Expressions. In N. Indurkhya and F.J. Damerau (Eds.)

[11] https://github.com/racai-ai/RoLLOD/blob/RRT-DCV/alignment.csv

*Handbook of Natural Language Processing*, Second Edition. Boca Raton: CRC Press, pp. 267–292.

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

Barbu, A.M. (2018). Dictionary of Verbal Contexts for the Romanian Language. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 17-21 July, Ljubljana, Ljubljana University Press, Faculty of Arts, pp. 411–422.

Barbu Mititelu, V. (2018). Modern Syntactic Analysis of Romanian. In O. Ichim, L. Botoşineanu, D. Butnaru, M.-R. Clim, V. Olariu (Eds.), *Clasic şi modern în cercetarea filologică românească actuală*. Iaşi: Editura Universităţii "Alexandru Ioan Cuza", pp. 67–78.

Barbu Mititelu, V. and Mitrofan, M. (2020). The Romanian Medical Treebank - SiMoNERo. In *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR-2020)*, pp. 7–16.

Barbu Mititelu, V., Cristescu, M., Nevaci, M. (2021). Un instrument modern de studiu al dialectului aromân: corpus adnotat morfosintactic. In M. Nevaci, I. Floarea, I.-M. Farcaş (Eds.), *Ex Oriente lux: In honorem Nicolae Saramandu*. Alessandria: Edizioni dell'Orso, pp. 141–160.

Bojar, O., Semecký, J., Benešová, V. (2005). VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics* 83:3–17.

Chiarcos, C., McCrae, J., Cimiano, J., Fellbaum, Ch. (2013). Towards Open Data for Linguistics: Linguistic Linked Data. In A. Oltramari, P. Vossen, L. Qin, E. Hovy (Eds.) *New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing*. Berlin, Heidelberg: Springer.

Cinková, S. (2006). From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006),* pages 2170–2175. Genova, Italy, May. European Language Resources Association (ELRA).

Fellbaum, Ch. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, Ch. J. and Atkins, B.T.S. (1992). Towards a frame-based lexicon: the semantics of risk and its neighbors. In A. Lehrer and E. F. Kittay (editors, *Prames, Fields and Contrasts*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 75–102.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, Vaxjo, Sweden, pp. 57–68.

Hinrichs, E., Telljohann, H. (2009). Constructing a Valence Lexicon for a Treebank of German. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, Groningen, The Netherlands, pp. 41–52.

Lopatková, M. (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon, PBML 79-80, pp. 37-60.

Lopatková, M., Bojar, O., Semecký, J., Benešová, V. and Žabokrtský, Z. (2005). Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*, pp. 99–106.

de Marneffe, M.-C., Manning, Ch., Nivre, J., Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Mărănduc, C., and Bobicev, V. (2017). Non Standard Treebank Romania – Republic of Moldova in the Universal Dependencies. In *Proceedings of Conference on Mathematical Foundations of Informatics (MFOI-2017)*, pages 111–116, Chisinau, Moldova, November.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, Ch.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Palmer, M., Kingsbury, P. and Gildea, D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31(1):71–106.

Urešová, Z., Štěpánek, J., Hajič, J., Panevová, J., Mikulová, M. (2014). PDT-Vallex: Czech Valency lexicon linked to treebanks. https://lindat.mff.cuni.cz/services/PDT-Vallex/, Feb 2014.

Woliński, M., Hajnicz, E. (2021). Składnica: a constituency treebank of Polish harmonised with the Walenty valency dictionary. *Language Resources and Evaluation* 55:209–239.