Universal and Language-specific Dependency Relations for Analysing Romanian

Verginica Barbu Mititelu
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
Romania
vergi@racai.ro

Cătălina Mărănduc
Faculty of Computer Science,
"Al. I Cuza" University,
Romania
catalina.maranduc@info.uaic.ro

Elena Irimia
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
Romania
elena@racai.ro

Abstract

This paper is meant as a brief description of the Romanian syntax within the dependency framework, more specifically within the Universal Dependency (UD) framework, and is the result of a volunteer activity of mapping two independently created Romanian dependency treebanks to the UD specifications. This mapping process is not trivial, as concessions have to be made and solutions need to be found for various language specific phenomena. We highlight the specific characteristics of the UD relations in Romanian and argument the need for other relations. If they have already been defined for (an)other language(s) in the UD project, we adopt them.

1 Introduction

The context of the work presented below is the creation of various language resources for Romanian. Throughout time, several resources have been created, which are available on the Meta-Share platform (http://ws.racai.ro:9191/). Nevertheless, the need for a syntactically annotated corpus was underlined in (Trandabăț et al., 2012). In the last years, two treebanks for Romanian were created. Although using different sets of relations, they both adopted the dependency grammar formalism and were created in complete awareness of each other.

Perez (2014) and Mărănduc and Perez (2015) reported on a treebank of (now) 5800 sentences, with 121 657 words and an average of 21 words per sentence. The sentences belong to all functional styles and cover different historical periods (the translated English FrameNet, Orwell's "1984", some Romanian belletristic texts, Wikipedia and Acquis Communautaire documents, political texts, etc.). They are annotated with dependency relations, but using a set of Romanian traditional grammar labels for the syntactic relations (such as prepositional attribute, adjectival attribute, direct complement, secondary complement, etc.). We refer to this corpus as UAIC-RoTb (the Romanian treebank created at "Al. I. Cuza" University of Iași).

Irimia and Barbu Mititelu (2015) report on a treebank (created at RACAI and further referred to as RACAI-RoTb) of (now) 5000 sentences. This corpus contains 5 sub-sections, covering the following genres: journalistic (news and editorials), pharmaceutical and medical short texts, legalese, biographies and critical reviews, fiction. From each such subsection of the Romanian balanced corpus (ROMBAC, Ion et al., 2012), the most frequent 500 verbs were selected and 2 sentences (with length varying from 10 to 30 words), illustrating the usage of each verb (so a total of 10 sentences per verb), were designated to be part of the treebank. They are annotated with dependency relations, but using a reduced set of labels, created with an eye to the UD set, but treating functional words as heads, differentiating among more types of objects (direct, indirect, secondary and prepositional) and disregarding the morpho-syntactic realizations of subjects and objects (so making no distinction between subjects or objects realized as nouns and subjects or objects realized as subordinate clauses, nor between subjects in active or in passive sentences).

Our effort now is to create a reference dependency Romanian treebank following the principles of the UD project by converting the annotation of these two treebanks into the UD style. The conversion process has not started yet, so we cannot report on any data about its performance. However, each team (the UAIC

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 28–37,
Uppsala, Sweden, August 24–26 2015.

and the RACAI one) has mapped the set of relations in their treebank to the UD set. For most of the situations, the two teams agree on the UD relations meant to describe various syntactic phenomena. However, there are cases when different solutions were given, as will be signalled below.

On the one hand, we will discuss below the UD relations from the perspective of their morpho-syntactic realization in Romanian, thus emphasizing language characteristics (section 3). On the other hand, we will describe language-specific constructions and bring arguments in favour of the treatment we propose (section 4). What we consider language-specific constructions are not necessarily constructions occurring only in Romanian. When they have been described for other languages as well, we will, in fact, add one more language argument supporting the respective relation.

2 Related work

Our effort of converting the treebanks in the UD annotation style is not singular. On the contrary, it aligns with the increasing number of such volunteer initiatives meant to offer treebanks for different languages consistently annotated, that could further help the development of multilingual parsers.

The 28 languages involved in this project now are Amharic, Ancient Greek, Basque, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, English, Finish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Latin, Japanese, Korean, Persian, Romanian, Slovenian, Spanish, and Sweden. We can notice the world wide interest for this topic, both for spoken and for dead languages.

The desideratum in the UD project is to have consistent annotations of treebanks for different languages. Consequently, all teams adopt the same relations for syntactic analysis. Nevertheless, language specific phenomena benefit of close attention and, besides the universal set of relations, extensions are also possible in order to accommodate all linguistic phenomena. For example, the Czech, English, Finnish, Greek, Irish, and Swedish teams have already proposed some extensions, for a correct annotation of the reflexive marker of passive voice (Czech), of the possessive nominal constructions (English, Finnish, Irish, Swedish), of relative clauses (English, Finnish, Greek, Irish, Swedish), etc.

3 Universal dependency relations in Romanian

Our intention of automatically converting the two treebanks (UAIC-RoTb and RACAI-RoTb) to the UD annotation style was motivated by the need for a bigger, unified, harmonious, conformant to international standards resource. In the conversion process, we confronted various problems connected to the representation of language phenomena within the new formalism. The way we decided to deal with them is described below.

For marking the syntactic relations between parts of speech in Romanian, we have used the inventory of relations from the UD project (http://universaldependencies.github.io/docs/u/dep/index.html, an adapted version of the relations described in de Marneffe, 2014):

Relation label	Description
root	the head of a sentence
nsubj	nominal subject
nsubjpass	passive nominal subject
csubj	clausal subject
csubjpass	clausal passive subject
dobj	direct object
iobj	indirect object
ccomp	clausal complement
xcomp	open clausal complement
nmod	nominal modifier
advmod	adverbial modifier
advcl	adverbial clause modifier
neg	negation
appos	apposition
amod	adjectival modifier
acl	clausal modifier of a noun (adjectival clause)
det	determiner
case	case marking
vocative	addressee
aux	auxiliary verb
auxpass	passive auxiliary
cop	copula verb
mark	subordinating conjunction
expl	expletive
conj	conjunct
cc	coordinating conjunction
discourse	discourse element
compound	relation for marking

	compound words
`name`	names
`mwe`	multiword expressions that are not names
`foreign`	text in a foreign language
`goeswith`	two parts of a word that are separated in text
`list`	used for chains of comparable elements
`dislocated`	dislocated elements
`parataxis`	parataxis
`remnant`	remnant in ellipsis
`reparandum`	overridden disfluency
`punct`	punctuation
`dep`	unspecified dependency

Table 1. UD relations used for annotating the Romanian treebank.

We do not use the `nummod` relation, as we treat numerals as either nouns or adjectives.
We will highlight below the specific characteristics of some of these relations in the analysis of Romanian and what decision regarding annotation they involved.

3.1. `Root`

In our treebank the predicate of a sentence can be a verb, an adverb (what Romanian traditional grammar calls a predicative adverb) (1, 2), an interjection (3), a noun (4) or an adjective (5). When such a predicate is the head of a sentence, it is marked as `root`. Although cases when an adverb or an interjection is the root of a sentence are not mentioned on the UD website, we consider them possible in sentences similar to the ones exemplified for Romanian.

(1) **Jos** mafia!
Down mafia!
“Down with the mafia!”

(2) **Poate** că întârzie.
Maybe that is_late
“He may be late.”

(3) **Marș** afară!
Shoo out!
“Get out!”

(4) Maria este **sora** mea.
Mary is sister-the my
“Mary is my sister.”

(5) Maria este **înaltă**.
“Mary is tall.”

If verbs, adverbs and interjections are commonly treated as predicates in Romanian linguistics, the last two are the result of adopting from UD the analysis of the copula *fi* “be” as being in `cop` relation with what traditional grammar analyses as a predicative.

Another situation when the root is not a predicate is represented by elliptical sentences, which lack a predicate, and thus their root is the head of the phrase they contain: in the Bi sentence below it is the noun *parc*. In case more than one argument or adjunct of the missing root are present, the head of the first one (in linear order) is the root of the sentence and all the others are attached to it by the relation they would have been attached to the verbal root if it had been present:

(6) A: Unde pleci?
Where leave-you?
“Where are you going?”
B: i) În **parc**.
In park
“To the park.”
ii) În **parc**, cu Dan.
In park, with Dan
“To the park, with Dan.”

3.2. `Cop`

In UD the copula *be* is linked by means of the relation *cop* to the predicative noun or adjective functioning as the root of the sentence. However, when the predicative is a clause, *be* is the root of the sentence and the clause predicative is `ccomp`. We adopted the same analysis for its Romanian equivalent, *fi*, in spite of the inconsistency in the analysis of this verb.

On the other hand, we can notice an inconsistent treatment of copular verbs in UD. Thus, the verb *be* is in `cop` relation to the root, whereas other copular verbs are analysed as roots: here is an example with *become* from the English treebank in its first release on the UD website (file en-ud-dev.conllu):

(7) John has **become** an engineer.
`root (become)`
`xcomp (become, engineer)`

In Romanian, the verb *deveni* “become” is always traditionally analysed as copular, whereas all the other copular verbs can also be predicative for some of their meanings. We illustrate this with *însemna*, which is predicative in (8a) and copular in (8b), according to the traditional grammar analysis:

(8) a) Copilul **a însemnat** tema.
Child-the has marked homework-the
"The child marked the homework."
b) Răspunsul lui **a însemnat** diplomație.
Answer-the his has meant diplomacy
"His answer meant/was_a_proof_of diplomacy."

In (8a) *tema* is the direct object and in (8b) *diplomație* is the predicative, not a direct object, as it does not pass the test specific to direct objects: substitution with an Accusative personal pronoun. Although the sentences may seem syntactically similar, they are different and traditional syntactic analysis captures the difference by assigning a distinct syntactic function to the two nouns following the verb.

Our solution for copular verbs (except *fi*, whose analysis is presented above), in line with other languages in the project, is to mark them as roots and treat them as regular raising verbs, so they take (i.e., their predicative is analysed as) an `xcomp` dependent. Consequently, the distinction between the two morphological values of such verbs (predicative and copular) is reflected in the different types of relation linking its second argument.

3.3. Subject

Subject is the only relation for which subtypes were created in UD in order to differentiate between active and passive sentences, on the one hand, and phrasal and clausal realization, on the other. Thus, four subtypes are used: `nsubj`, `nsubjpass`, `csubj`, `csubjpass`, which we adopted.

In Romanian, the nominal subject is sometimes doubled by a pronominal one, marking a certain illocutionary attitude of the speaker: threat, promise, and reassurance (see 9). As Romanian is a pro-drop language, the nominal subject may be omitted (10). Irrespective of the presence or absence of the nominal subject, the pronoun has a clitic behaviour in such examples (Barbu, 2003).

The analysis we propose within UD is the following: the nominal, when present, is marked as `nsubj`, while the pronoun in Nominative case is marked as `expl`, with *și* as `advmod`. The analysis of the pronominal doubling subject does not depend on the presence or absence of the nominal subject.

(9) Tata vine și **el** imediat.
Father-the comes and he immediately
"Father will also come immediately."

(10) Vine și **el** imediat.
Comes and he immediately
"He will also come immediately."

3.4. Objects

Direct, indirect, secondary objects. The Grammar of Romanian Language (GRL) describes three types of objects: direct, indirect and secondary. The last one is an object in the Accusative case, co-occurring with a direct object, also in Accusative. When only one Accusative object occurs with a verb, that object is always a direct one (see 12b). While the direct object may co-occur with either the indirect or the secondary object, the other two can never co-occur:

(11) Fata a dat nume păpușilor.
Girl-the has given names dolls-the-to
"The girl gave names to the dolls."

(12) a) Bunica i-a învățat pe copii o poezie.
Grandmother-the them-has taught PE children a poem
"Grandmother taught the children a poem."
b) Bunica a învățat o poezie.
Grandmother-the has learned a poem
"Grandmother has learned a poem."

Within UD, we analyse the direct object in (11) (*nume*) as `dobj` and the indirect object (*păpușilor*) as `iobj`. As in UD there is no label for the secondary object, in (12a) the direct object (*copii*) is analyzed as `iobj` and the secondary object (*poezie*) as `dobj`, adopting the Czech convention, supported by the semantic roles distribution in the sentence: the animate object is the addressee, and the non-animate is the patient.

Thus, unlike traditional grammar, when it is not the only object of the verb, the Accusative object is either direct or indirect, depending on the co-occurring object: when there is a Dative and an Accusative object, the Dative is `iobj`, and the Accusative is `dobj`; when two Accusatives co-occur, the [+Animate] one is `iobj`, and the [-Animate] one is `dobj`. So, an automatic analysis needs access to a word sense disambiguation tool or to a dictionary.

Object doubling. A characteristic of Romanian direct and indirect objects is their obligatory doubling by a clitic, when certain characteristics hold: for the direct object: definiteness, pre-verbal occurrence, co-occurrence with the preposition *pe*, pronominal realization; for the indirect object: [+Human], preverbal occurrence.

Thus, the direct object can have the types of realizations presented under (13), while the indirect object those under (14):

(13) a) Ascult **muzică**.
Listen-I music.
"I am listening to music."
b) **Îl** ascult pe **Ion/el**.
Cl.3.sg.masc.Acc. listen-I PE John/him.
"I am listening to John/him."
c) **Îl** ascult.
Him listen-I
"I am listening to him."
(14) a) Dau de mâncare **pisicii**.
Give-I of food cat-the-to
"I give food to the cat."
b) **Le** dau de mâncare **copiilor**/**lor**.
Cl.3.pl.Dat. give-I of food children-the-to/to-them
"I give the children/them food."
c) **Le** dau de mâncare.
To-them give-I of food
"I give them food."

When the direct or indirect object is not doubled, it is analysed as `dobj` and `iobj`, respectively, no matter if it is realised by a noun or a pronoun (see examples a) and c) under (13) and (14)). In the b) examples, the clitic is analysed as `expl` and it doubles a `dobj` or `iobj`, respectively.

3.5. Adverb modifiers

Adverbs can modify nouns (15), verbs (16), adjectives (17) and other adverbs (18) in Romanian and for all these cases we use the label `advmod`.

(15) Cititul **noaptea** nu este sănătos.
Reading-the at-night not is healthy
"Reading at night is not healthy."
(16) Citesc **noaptea**.
Read-I at-night
"I read at night."
(17) o casă **chiar** frumoasă
a house really beautiful
"a really beautiful house"
(18) Scrie **chiar** ordonat.
Writes really neatly
"He writes really neatly."

However, with some verbs, the adverb represents an obligatory dependent, without which the sentence is ungrammatical:

(19) Copilul se poartă *(**frumos**).
Child-the refl.cl.3.sg. behaves beautifully
"The child behaves himself."

As a consequence, in Romanian we use the `advmod` label both for non-core dependents and for core ones.

3.6. Subordinate clauses

Subordinate clauses are introduced by relative elements (and indefinites formed from relatives) or subordinating conjunctions. The relative elements are pronouns, adjectives or adverbs. The major difference between relatives (and indefinites) and conjunctions concerns their syntactic role within the clause they introduce: the former have a syntactic function in the subordinated clause, whereas the conjunctions lack it. As a consequence, we adopted the UD solution of treating them in different ways: relatives (and indefinites) establish a relation of whatever kind (`nsubj`, `dobj`, `iobj`, `advmod`, `amod`, etc.) with the head of the subordinated clause (20); the subordinating conjunction is only a marker of the syntactic subordination and establishes the relation `mark` with the head of the subordinated clause (21).

(20) Știu **cine** a venit.
Know-I who has come
"I know who has come."
`nsubj(venit, cine)`
`ccomp(Știu, venit)`
(21) Știu **că** vine târziu.
Know-I that comes late
"I know that (s)he comes late."
`mark(vine, că)`
`ccomp(Știu, vine)`

This way, we ensure, in fact, a consistent way of choosing the element in the subordinated clause meant to participate to the subordinating relation: the head of the subordinate clause.

A consistent annotation is ensured also for the relative elements, which can also function as interrogative elements in questions: they always establish a syntactic relation with the head of the clause:

(22) **Cine** a venit?
"Who has come?"

The conjunctive mood is formed with the conjunction *să*. It can occur both in main clauses (23) and in subordinate ones (24).

(23) **Să** mergem!
SĂ go-we
"Let's go!"
(24) Vreau **să** mergem.
Want-I SĂ go-we.
"I want us to go."

Our solution is to analyse both such occurrences in the same way, i.e. *să* is `mark` for the verb, in spite of the UD definition of the marker as a "word introducing a finite clause subordinate to another clause" (cf. http://universaldependencies.github.io/docs/u/dep/mark.html).

4 Language-specific constructions

In this section we describe constructions from Romanian for which the UD relations are not appropriate.

4.1. Agent complement

An agent complement may occur in constructions with the verb in the passive voice (25) or with non-finite verbs (26) or adjectives (27) with a passive meaning:

(25) Cartea a fost cumpărată **de Ion**.
Book-the has been bought by John
"The book was bought by John."
(26) Aceasta este calea de urmat **de_către** orice **om** integru.
This is way-the of followed by any man honest
"This is the way to follow for any honest man."
(27) Avea un comportament inacceptabil **de_către colegii** săi.
Had-he a behaviour unacceptable by colleagues-the his
"He had an unacceptable behaviour by his colleagues."

Besides the prepositional phrase (headed by the simple preposition *de* or by the compound preposition *de_către*[1]), the agent complement may also be realized by a subordinate relative clause:

(28) A fost angajat **de cine a avut încredere în el**.
Has been hired by who has had trust in him.
"He was hired by who trusted him."

In line with other languages displaying this syntactic specificity in the UD project (Swedish), we support the proposal of creating a subtype of the `nmod` relation: `nmod:agent`. We highlight the fact that in such cases `nmod` is also a core dependent of the head. For the last example, when the agent is realized as a subordinate clause (28), we propose `ccomp:agent`.

4.2. Prepositional object

This is a verb argument (i.e., it is part of the verb subcategorization frame) introduced by a preposition selected by the verb:

(29) Mă gândesc **la Maria**.
Refl.cl.1.sg.Acc. think of Mary
"I am thinking of Mary."

Prepositions are not heads in UD. So, the nominal is annotated as `nmod` on the verb and the preposition as `case` on the noun. However, `nmods` are defined as non-core dependents of a predicate in UD. Thus, annotating the prepositional objects as `nmod` implies treating them in exactly the same way as we treat adverbials realized by a prepositional phrase. In the following example, *la problemă* is the prepositional object and *la masa* is the time adverbial, in traditional grammar terms.

(30) Mă gândesc **la problemă la masa** de prânz.
Refl.cl.1.sg.Acc. think of problem at meal-the of noon
"I am thinking at the problem at lunch."

However, if `nmods` functioning as adverbials are optional, prepositional objects are obligato-

[1] In the pre-processing phase, compound prepositions are recognised (given their presence in our electronic lexicon) and marked as one token (using the underscore).

ry for the grammatical correctness of the sentence:

(31) Mă bazez *(**pe voi**).
Refl.cl.1.sg.Acc. count-I *(on you)
"I count *(on you)."

That is why we are not satisfied with this analysis of prepositional objects in which they are not distinguished from dependents which are not obligatory and we propose to redefine the `nmod` relation so that it covers both core and non-core dependents. In line with this redefinition, in RACAI-RoTb we introduce the `nmod:pmod` subtype of `nmod` to account for the obligatory prepositional objects of predicates, a phenomenon present in other languages, as well. However, in UAIC-RoTb such cases are analysed as `iobj`, given the occurrence in language of two parallel structures for indirect object: one with the noun in Dative case and another with the preposition *la* and the noun in Accusative. The latter structure is the norm for phrases containing a quantifier or a numeral in the standard language (32), but it witnesses an extension to all kinds of nouns in colloquial speech (33):

(32) Le spun o poveste *la trei copii*.
"I tell a story to three children."
(33) Le spun o poveste *la copii*.
"I tell a story to the children."

4.3. Possession

There are several ways of expressing possession in Romanian: sentences with the verb *avea* "to have" or its synonyms, genitive nouns or personal pronouns, possessive adjective (which we link by means of the `amod:poss` relation to the head nominal, see (4) above, where *mea* is in `amod:poss` relation with its head, *sora*) and pronouns and dative personal pronouns. We focus here on genitive and dative constructions, as the others do not raise any special problems.

The genitive constructions (involving nouns or personal pronouns) may have a possessive meaning (34) or not (35):

(34) Trecutul **castelului** este necunoscut.
Past-the of-castle-the is unknown
"The past of the castle is unknown."
(35) Reconstrucția **castelului** a început.
Rebuilding of-castle-the has started
"The rebuilding of the castle has started."

And this is the case in other languages as well: see Finish (http://universaldependencies.github.io/docs/fi/dep/nmod-poss.html, accessed on April 7). The subtype `nmod:poss` is used to annotate all these constructions, in spite of the semantic differences between them. And this is the way in which such cases are dealt with in UAIC-RoTb, as well. However, the RACAI-RoTB team uses only the label `nmod`, leaving the possessive value of genitives not specified.

As far as the possessive dative is concerned, it is always realised by a pronominal clitic on the verb:

(36) **Mi**-am pierdut fularul (*meu).
Cl.1.sg.Dat-have-I lost scraf-the (*my)
"I have lost my scarf."

The co-occurrence of the possessive adjective (*meu*) in such constructions makes them pleonastic.

For the clitic analysis the RACAI-RoTb team decided to use the `nmod:poss` relation to link it to the verb. The UAIC-RoTb team opted for the `iobj` relation for such cases.

4.4. Reflexive pronouns

Reflexive pronouns can have various semantic values:

- reflexive value: see examples (29), (30) and (31) above;
- reciprocal value:

(37) Doi copii **se** bat.
Two children SE fight
"Two children are fighting."

- passive value:

(38) **Se** bat albușurile cu zahăr.
SE beat whites with sugar
"Egg whites are beaten with sugar."

- pronominal value:

(39) Ion **se** spală.
John SE washes
"John is washing himself."

- impersonal value:

(40) **Se** înnoptează.
SE gets_dark
"It is getting dark."

For the reflexive, reciprocal and impersonal value, when the reflexive pronoun (either in Accusative or in Dative case) has no syntactic function and is a mere marker of the reflexive, reciprocal or impersonal voice of the verb, according to traditional grammar, we adopt the relation `compound:reflex`, a subtype of the `compound` relation, to link the pronoun to the verb, as proposed for Czech.

For the passive value, when the occurrence of the pronoun blocks the occurrence of the passive auxiliary (*fi*), we propose the relation `auxpass:reflex`, a subtype of the `auxpass` relation, to link the pronoun to the verb.

For the pronominal value, we need no other relation, as the pronoun has a syntactic function: `dobj` or `iobj` (in (37) it is a `dobj`).

4.5. Participles

The Romanian participle has some characteristics that make it similar to adjectives (it inflects for number, for gender and for case and can modify a noun) and others that prove its verbal nature (it can take arguments):

(41) poezii **recitate** de meseni la comanda lui Charles
poems recited by diners at order-the def.art.masc.sg.Genit. Charles
"poems recited by diners at Charles' order"

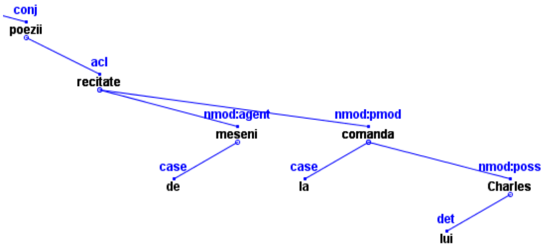

Fig. 1. The arguments of the participle *recitate*.

Given the participle possibility of having arguments, we decided to analyse the participles that determine a noun as establishing the `acl` relation to that noun.

4.6. Putting semantics into adverbials

UAIC-RoTb contains semantic information about the adjuncts occurring therein: they express time, place, manner, instrument, exception, purpose, cause, etc. They are morphologically realised as adverbs, noun phrases, prepositional phrases (containing a noun) or subordinate clauses. Considering potential further processing of the treebank for various applications, a part of the semantic information was preserved, namely the time adjuncts. They are annotated as `advmod:time`, `nmod:time` or `advcl:time`, respectively.

4.7. Infinitive or conjunctive?

A specific syntactic feature is the verb mood selected for expressing the clausal argument of a verb. UAIC-RoTb has an incipient parallel treebank containing 250 sentences of the novel "1984" by G. Orwell, annotated in English, French and Romanian, which allows us to compare the syntax of the three languages. In English and in French the second verb is an infinitive directly related to the first one or related by means of a preposition:

(42) Il cesse **de** parler / He ceases **to** speak / El încetează **să** vorbească.

In Romanian the conjunctive mood is selected, which has the conjunction *să* as a marker. The structure with the second verb in the infinitive with preposition is possible in Romanian but less frequent and either obsolete or formal.

(43) Noi încetăm (de) **a vorbi**.

The Romanian subjunctive has inflexion for person and number:

(44) Nous cessons de parler. / We cease to speak. / Noi încetăm să **vorbim.**

Thus, in Romanian we can have either two clauses (when the second verb is in the conjunctive mood) or only one (when the second verb is in the infinitive mood), in traditional grammar terms. Both cases correspond to English and French structures with a non-finite verb. However, this issue disappears as the dependency grammar treats all verbs identically, i.e. as heads of clauses, irrespective of their finite or non-finite form.

4.8. The verb *a putea* "can"

The problem of the mood of the second verb in Romanian gets more complicated if we compare the structures containing modal verbs in the three languages.

(45) We **must** eat. /Il **faut** manger. /**Trebuie** să mâncăm.

In the languages that have modal verbs, they take short infinitive. In Romanian, among the potential modal verbs, only *a putea* "can" displays this syntactic behaviour, as well as the usual one, with the second verb at the subjunctive mood.

(46) Putem **scrie.** / Putem **să scriem.**
"We can write".

Romanian does not have modal verbs. However, there are a number of syntactic phenomena that make us conclude that *a putea* is the only verb in the process of transition to the status of modal verb.

The constructions with the verb *a putea* followed by a short infinitive are synonymous and commutable with those where it is followed by a conjunctive (see 46). Statistically, the infinitive is more frequent than the conjunctive: out of 150 examples containing this verb in UAIC-RoTb, 33% contain a conjunctive, 24% contain no following verb (so, they are statistically irrelevant), and 43% contain a short infinitive without any preposition.

There are a lot of dependents of the verb *a putea* that are advanced one level up in the tree: originally, they are arguments of the infinitive verb occurring after *a putea*:

(47) Problema țărănească nu se poate rezolva.
Problem-the rustic not SE can solve
"The peasants' problem cannot be solved".

The subject *problema* belongs to the subcategorization frame of the verb *rezolva*. However, its number agreement with the verb *poate* proves its new syntactic status, that of subject of *poate*. *Se* is the passive maker of the verb *rezolva*, although raised on *poate*.

Other core-dependents are also raised on the verb *a putea*: here is an example with an indirect object:

(48) Nu-**mi** putea da o cameră.
Not-to-me could-he give a room
"He could not give me a room."

We consider that *a putea* should be analysed as `aux` when followed by an infinitive, and as a root when followed by a subjunctive.

5 Conclusion

The Universal Dependency grammar project offers the material for a comparative and contrastive study of the languages involved in it. The same phenomenon can be studied in various languages and similarities, as well as differences highlighted.

During our process of automatically converting the annotation of the two Romanian treebanks into UD annotation, we had to find solutions for various language phenomena and they were either of the type "use a UD label to cover more situations than those presented within the UD project" or of the type "postulate a new label, a subtype of a relation existing in UD".

One of the results of our working methodology is the heterogeneity of the syntactic relations covered by a UD label: see the case of `nmod` presented above. Another result is the blurring of the very clear border between some syntactic functions: see the case of direct object, indirect object and secondary object.

References

Blanca Arias, Núria Bel, Mercè Lorente, Montserrat Marimón, Alba Milà, Jorge Vivaldi, Muntsa Padró, Marina Fomicheva, Imanol Larrea. 2014. Boosting the Creation of a Treebank. In Calzolari, Nicoletta, Choukri, Khalid; Declerck, Thierry (et al.) (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14): May 26-31, 2014 Reykjavik, Iceland*. [s.l.]: ELRA. p. 775-781.

Verginica Barbu. 2003. Construcții cu subiect dublu în limba română actuală. O perspectivă HPSG. In G. Pană Dindelegan, *Aspecte ale dinamicii limbii române actuale*. Editura Universității din București, p. 73-79.

GRL – V. Guțu Romalo (ed.). 2005. *The Grammar of Romanian Language*. Romanian Academy Publishing House, second volume.

Radu Ion, Elena Irimia, Dan Ștefănescu, Dan Tufiș 2012. ROMBAC: The Romanian Balanced Annotated Corpus. In *Proc. LREC'12* Istanbul, Turkey.

Elena Irimia and Verginica Barbu Mititelu. 2015. *Building a Romanian Dependency Treebank*. Corpus Linguistics 2015, Lancaster, UK, 21-24 July 2015.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gül¸sen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A languageindependent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135.

Montserrat Marimon and Nuria Bel. 2014. Dependency structure annotation in the IULA Spanish LSP Treebank. *Language Resources and Evaluation*. Amsterdam: Springer Netherlands.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-lingustic typology, *Proceedings of LREC 2014*: 4585-4592.

Cătălina Mărănduc and Augusto-Cenel Perez. 2015. *A Romanian dependency treebank*, CICLing 2015, Cairo, 14-20 April.

Augusto-Cenel Perez. 2014. *Resurse lingvistice pentru prelucrarea limbajului natural*. PhD thesis, "Al. I Cuza" University, Iasi.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik. 1985. A Comprehensive Grammar of the English Language. Longman.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale.* Klincksieck, Paris.

Diana Trandabăţ, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, Dan Tufiş. 2012. *The Romanian Language in the Digital Age. Limba română în era digitală*. In White Papers Series (Rehm, Georg and Uszkoreit, Hans). Springer-Verlag, Berlin, Heidelberg.