Modern Syntactic Analysis of Romanian

Verginica Barbu Mititelu[*]

Keywords: *syntax*; *Romanian*; *Treebank*; *dependency grammar*; *universal dependencies*

1. Introduction

The development of language resources has become a popular endeavour among computational linguists and corpus linguists, especially in the last decades. The motivation behind this is, in fact, twofold: on the one hand, language resources are interesting from a linguistic point of view: they may apply a certain grammatical theory or, even if theory neutral, they reflect the language and can serve as a base of linguistic knowledge for dictionaries, grammars, etc.; on the other hand, such resources can be used for training and testing automatic tools for processing languages.

A corpus (i.e., a collection of electronic texts) annotated at the syntactic level is called a treebank. It reflects the syntactic groups within a sentence, the relations between them, as well as the relations between their components. As syntactic annotation always comes on top of previous morphologic annotation, the morphologic realizations of various syntactic functions or of the linguistic units entering a certain relation are also explicit in treebanks.

In this paper we present a treebank for Romanian, containing texts from various language registers, annotated according to the principles and the set of relations in the Universal Dependencies project (universaldependencies.org). We also show how it is accessible and how it can be queried.

2. The treebank structure

The corpus that makes the treebank contains 9523 sentences. They are not consecutive sentences extracted from texts. They were extracted as separate units, so as to serve several aims: coverage of various functional styles (journalistic, legal, scientific, imaginative), coverage of various domains (medicine, computer science, mathematics, literary theory, law, etc.), different sentence lengths and different authors, different types of sentences (declarative, interrogative, imperative and exclamative, on the one hand, and simple, compound, complex and complex-compound ones, on the other hand). As far as the

[*] "Mihai Drăgănescu" Romanian Academy Research Institute for Artificial Intelligence (ICIA), Bucharest, Romania.

originality of the sentences is concerned, two aspects are important: firstly, the vast majority of them are original creation, they are not translations; secondly, no intervention occurred in the sentences. However, sentences distribution according to the previously mentioned categories is not even. We illustrate here with the functional styles. In the table below, for each style we present the number of sentences it is represented by, their distribution in the whole corpus with respect to the number of sentences contained, the total number of tokens in these sentences, their distribution in the whole corpus with respect to the number of tokens contained, and the average sentence length.

Table 1. Corpus structure according to the functional style

Style	No. of sentences	Distribution in the whole corpus by the number of sentences	No. of tokens	Distribution in the whole corpus by the number of tokens	Average sentence length
imaginative	1804	19%	18940	9%	10
journalistic	932	10%	23345	12%	25
legal	1520	16%	45620	23%	30
scientific	2903	30%	65188	32%	22
miscellanea	2364	25%	47186	24%	20
TOTAL	**9523**	**100%**	**200248**	**100%**	

Sentence-wise, almost a third of the corpus is represented by the scientific style. A quarter of it is a mixture of approximately the same number of sentences from all styles, which constituted the original core of the treebank and was kept separate in order to help the automatic processing of the sentences. The next best represented style is the imaginative one, followed by the legal one. The journalistic one is the least well represented. However, token-wise, the distribution of the styles in the whole corpus is not the same, given the average sentence length in each style, a measure which shows huge variation: the legal sentences are the longest ones (with an average length of 30 tokens per sentence), whereas the imaginative ones are the shortest, their average length being of only 10 tokens per sentence. Consequently, the scientific style remains the best represented, followed (by the miscellanea collection and then) by the legal and then the journalistic one. The imaginative subcorpus is, in fact, the least well represented, from this perspective.

3. Corpus processing and annotation levels

All sentences in the corpus were first tokenised: words and punctuation were identified. Although this may seem a trivial task, it raises several problems: words are not always separated by blanks: consider the string *n-am citit*

(not-have_I read "I haven't read"): three words must be separated here: *n-*, *am*, *citit*. One must note that the hyphen is not considered punctuation, thus it is not identified as a token here. Moreover, there are cases when the hyphen goes with the first token (as in the previous example) and other cases when it goes with the second one, as in the string *am citit-o* (have_I read-it "I have read it"), in which the tokens are *am*, *citit*, *-o*. Although punctuation is usually a separate token, this is not always so: consider the case of *etc.*, where the whole string (so the dot included) is one single token, so the dot is not a token in such cases. These are only a few problems that must be dealt with in the process of tokenisation.

After being tokenised, the sentences are part of speech tagged, a process which consists in the identification of the part of speech for each word, as well as of the values of its morphological categories (which we call attributes): for example, for nouns the following attributes apply: type (with the possible values common or proper), gender (masculine and feminine), number (singular or plural), case (direct, oblique or vocative) and definiteness (definite or indefinite). The parts of speech and their attributes were defined in a multilingual context, in the project MULTEXT-East (Erjavec 2012).

The next step was to lemmatise the words, that is to specify their lemma.

Tokenisation, part of speech tagging and lemmatisation were done automatically, with no manual intervention, with an in-house tool called TTL (Ion 2007), whose accuracy is 97.5% (Tufiş *et al.* 2008). The result was the creation of an annotated corpus (different from a treebank). The syntactic annotation, which is what makes a corpus become a treebank, was a semi-automatic process, involving both manual and automatic annotation.

The treebank we describe here is called RoRefTrees. It was meant to be a reference treebank for Romanian, on which a syntactic parser to be trained and tested. Moreover, its development followed current trends in treebanks annotation, namely the Universal Dependencies (UD) project, aiming at offering a set of principles and of syntactic relations as general as to be applicable to all languages. At the moment of RoRefTrees creation, version 1.4 of the UD guidelines were observed.

RoRefTrees is based on two existing and accessible treebanks: UAIC-RoDepTb (Perez 2014) and RACAI-RoTb (Irimia, Barbu Mititelu 2015). This means that the sentences were chosen from these treebanks. However, these treebanks differ in their annotation at all levels. That is why, all selected sentences were subject to re-annotation with the TTL tool, as presented above. For syntax, though, a different procedure was followed.

Syntactically, there are both similarities and differences on the one hand between UAIC-RoDepTb and RACAI-RoTb (as described in Barbu Mititelu *et al.* 2016) and on the other hand between these two treebanks and the UD principles and labels. A first step was to create a small treebank annotated

according to UD on which to train a parser (namely MALT parser, Nivre *et al.* 2007) so that to create a tool for automatic annotation of the rest of the corpus. We adopted this solution instead of the complete manual syntactic annotation of RoRefTrees because it involves less effort. This small corpus is the miscellanea component of RoRefTrees. All its sentences were manually annotated according to UD v. 1.4. After that, MALT parser was trained on it and then run on a pool of sentences, which were afterwards manually corrected. The process was iterative, in the sense that each new manually corrected set of sentences was added to the initial miscellanea set and the parser was retrained on this new set before being run on another set of sentences, thus obtaining better results at each run.

4. The principles of syntactic annotation in RoRefTrees

The dependency grammar is the background for the syntactic annotation in RoRefTrees. This formalism allows for the identification of head – dependent pairs and for their labelling. As opposed to constituency grammar, no phrases are identified in the dependency grammar.

The syntactic analysis of a sentence can be represented as a tree. All trees are rooted and their branches link two nodes in which tokens occur (so, punctuation is also included in the tree). One node can be the head of any number of other nodes but can have only one head, the exception being the root node which has no head. A node with no dependents is called a leaf. Punctuation can only occur in leaf nodes.

The root of this tree is the word which carries the main predication of the sentence and this is usually a verb. In cases of verb ellipsis, another word is chosen as the root, namely the first one, in linear order, on which other words depend.

For establishing the status of a word in a dependency relation, i.e. head or dependent, the following principles were observed:

(P1) Only content words can be heads. Thus, subordinate clauses are not headed by their subordinating conjunction, but by their verb.

(P2) Non-finite verb forms are considered heads of subordinate clauses.

(P3) The copula verb *a fi* ("to be") is a dependent.

(P4) In a structure with coordination, the head is the first conjunct; the rest of the conjuncts, the coordinating conjunction and the associated punctuation are all dependents of the first conjunct. This is a flat representation of structure with coordination (as opposed to the hierarchical representation).

(P5) A predicate can have more arguments, but they must be of a different type. This is of extreme importance for Romanian, which displays the doubling clitic phenomenon: the clitics will establish a different relation with the verb than the nominal (see below).

5. The set of relations in RoRefTrees

Working within UD also implies conforming to the set of relations established within this project. As already mentioned, UD aims at universality, that is, in this case, a set of syntactic relations applicable to all languages. However, whenever for a language there is a need for coining a new relation, this is possible, but only as a subtype of a universal one. Given the high number of languages in the project, a certain subtype is used for several languages.

The inventory of relations used for the analysis of Romanian is presented in Table 2 below. All relations written in boldface and preceded by an arrow are language-specific ones. Those that are used only for Romanian are both boldfaced and italicised, whereas those that are used for other languages as well, but are not universal are only boldfaced.

Table 2. The inventory of relations used for analysing Romanian

Core dependents of clausal predicates			**Non-core dependents of clausal predicates**			**Special clausal dependents**		
Nominal dep	*Predicate dep*		*Nominal dep*	*Predicate dep*	*Modifier word*	*Nominal dep*	*Auxiliary*	*Other*
nsubj	csubj		nmod	advcl	advmod	vocative	aux	mark
nsubjpass	csubjpass		***↳nmod:pmod***	***↳advcl:tcl***	***↳advmod:tmod***	discourse	auxpass	punct
dobj	ccomp	xcomp	**↳nmod:tmod**		neg	expl	cop	
iobj	***↳ccomp:pmod***		**↳nmod:agent**			**↳expl:pv**		
						↳expl:pass		
						↳expl:impers		
						↳expl:poss		
Noun dependents			**Compounding and unanalyzed**			**Coordination**		
Nominal dep	*Predicate dep*	*Modifier word*	compound	mwe		conj	cc	punct
nummod	acl	amod	name	foreign	goeswith		**↳cc:preconj**	
appos		det						
nmod		neg						
Case-marking, prepositions, possessive			**Loose joining relations**			**Other**		
case			list	parataxis	remnant	*Sentence head*	*Unspecified dependency*	
			dislocated		reparandum	root	dep	

One can notice relations in which the head is a predicate (see, in the table, core dependents of clausal predicates, non-core dependents of clausal predicates and special clausal dependents), relations in which the head is a noun (see noun dependents), relations between words in compound and unanalysable units (see compounding and unanalysed), relations used for coordination, for linking prepositions to their head (see case-marking, prepositions, possessive), and loose joining relations. The relation "root" and "dep" have a special status: the former is a pseudo-relation, as it links the root of the tree to an artificial node in the formal representation of the sentence; the latter is used when the annotator is unable to identify the type of a relation.

The table also shows that different labels are used for the relations linking a nominal or a clausal dependent to its head, even when their syntactic function is the same: e.g., see the existence of nsubj (nominal subject) and csubj (clausal subject).

One can notice in the above table that the core dependents of a predicate are the subject and the objects. For the former, there are four types, depending on their realisation (nominal or clausal) and also on the diathesis of the verb (nsubj and csubj are for verbs at the active voice, while nsubjpass and csubjpass are for verbs that are at the passive voice). This correlates with two labels for the auxiliaries: aux and, respectively, auxpass, the latter being used for the passive auxiliary.

The nominal objects are the direct object (labelled as dobj) and the indirect one (iobj). In the case of secondary objects, the annotation decision was to assign the dobj relation to the non-animated one and the iobj relation to the animate one (the assumption being that the animate object is semantically a beneficiary, which is often the case of indirect object as well). However, their clausal realisation is labelled identically: ccomp. This implies that when someone is interested in retrieving the clausal realisations of direct or indirect object in the treebank, they will have to manually distinguish among them. An exception to principle (P5) above, which prevents the existence of two relations of the same type on the same head, is the case when both the direct and the indirect object are realized as clauses and have the same label (namely ccomp).

For Romanian we introduced a subtype of this relation (ccomp:pmod) which links the head of the clausal realisation of a prepositional object to its head. The nominal realisation of this object is labelled as nmod:pmod, a subtype of the nmod relation, which links a nominal dependent (which is not a core one) to the predicate. Other subtypes of this relation are nmod:agent (linking the agent to its head) and nmod:tmod (linking the time complement to its head).

The clitics doubling a direct or indirect object (as well as the pronoun doubling the subject) are in expl relation with the verb. Some clitics may have other semantic meaning and, consequently, some subtypes were proposed for annotating them: expl:pv is for the reflexive clitic, expl:pass for the reflexive passive clitic, expl:impers for the impersonal and reciprocal value of the clitic and expl:poss for the Dative clitic with a possessive value.

Copula verbs have an inconsistent annotation throughout the treebank: the relation cop is used exclusively for linking the verb *a fi* ("to be") to its head (the predicative), thus the copula verb remains a leaf in the tree. There is only one exception to this treatment of the copula verb *a fi*, namely when its predicative is a clause: in such cases the copula verb is the head of the structure, so not a leaf in the tree. All the other copula verbs (e.g., *a deveni* ("to become"), *a se face* ("to become"), *a ieşi* ("to become"), etc.) are heads of their clauses. The predicative is linked by means of the xcomp relation to the verb. Besides this, xcomp also links secondary predicates to their head.

All adverbs are linked by the relation advmod to their verb, noun, adjective or adverb head. Only for the time adverb did we introduce a special label: advmod:tmod.

All adverbial clauses, irrespective of their type, establish the advcl relation with their head. We created a subtype of it (advcl:tcl) only for the time adverbial.

The dependents of a noun are nouns (and they are linked by means of the nmod relation), numerals (linked by means of the nummod relation), adjectives (linked by means of the amod relations), clauses (linked by means of the acl relation), appositions (appos) and determiners (det).

Prepositions are attached to their heads by means of the case relation. Coordinating conjunctions are the dependent in a cc relation, while the subordinate ones are linked by the relation mark to the head of the clause. The subtype cc:preconj is used for the correlative conjunction in pairs such as *fie…, fie…*

The negation is attached to a verb (sentential negation) or any other word (when expressing component negation).

The vocative relation needs no explanation. The relation discourse is used for interjections, fillers (such as *ăăă*) and several discourse markers. In Romanian compound is used only for numerals. The relation between the components of a locution and its head is mwe. A flat representation is given to names made up of at least two words, in which the first one in linear order is the head and the others are its dependents linked by means of the relation name. A string of foreign words is analysed in a similar way, but the relations is called foreign.

The relation remnant is used to link words from different clauses but with the same syntactic function, when the head of one of them is missing (so in elliptical clauses). In Figure 1 there are two such relations: the sentence consists of two clauses, coordinated, with a parallel syntactic structure and with a verbal ellipsis in the second one. The first relation remnant links the word "Maria" to its (artificial) head "Dan". This is possible because the sentences have the same syntactic structure (subject + predicate + direct object) and the two words occupy the same syntactic position (the subject). The explanation is similar for the second relation remnant, which links the direct object of the elliptical clause to the direct object of the first (non-elliptical) clause.

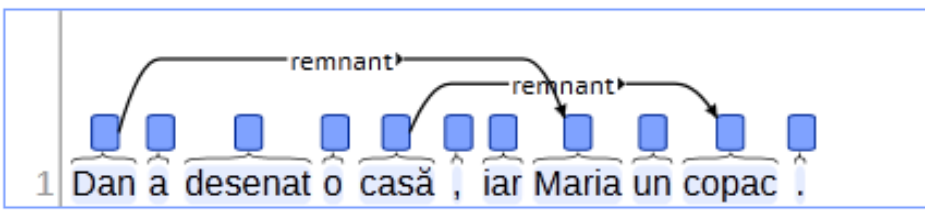

Figure 1. The relation remnant

Parataxis is the relation used for words and elements that are placed side by side, but without any explicit coordination or subordination between them.

Punctuation at the end of the sentence is attached to its root. However, punctuation which serves other roles is attached to other elements: for example, punctuation involved in enumeration (a kind of coordination) is attached to the first conjunct, while punctuation used with appositions is attached to the head of the apposition.

The other relations serve the need to annotate everything that may occur in a sentence, even if they are not grammatically relevant: e.g., the relation goes with is used for coping with cases when a word is accidentally broken into two parts by a blank: its second part is linked to the first one by this relation, whereas the first part will establish with its head the adequate syntactic relation.

In Figure 2 we show a tree representation of one of the sentences in RoRefTrees, which is written on the yellow line at the top of the figure. This shows the analysis of coordination, of a passive structure with a reflexive clitic, and of a prepositional object.

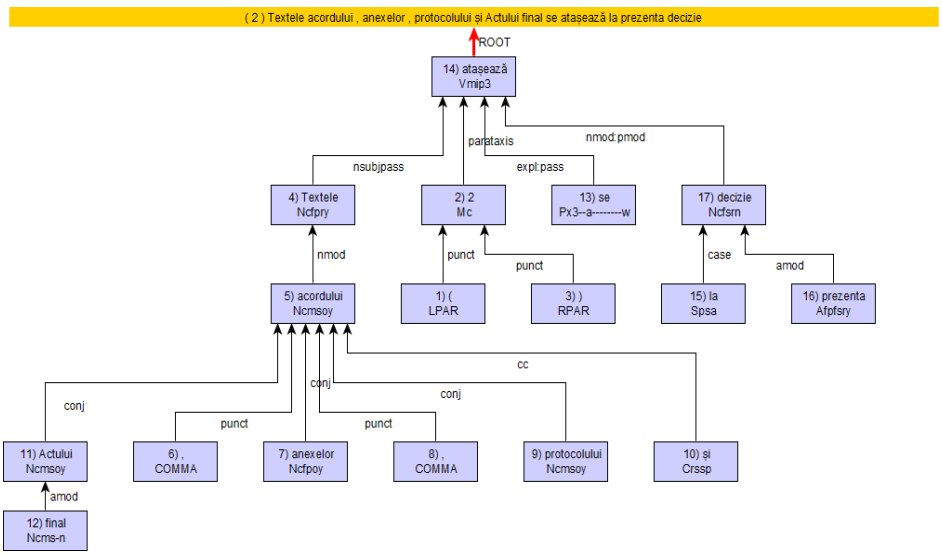

Figure 2. A tree representation of a sentence in RoRefTrees

6. Access to RoRefTrees

The treebank is publicly available for both download and query. It is officially released within the UD project, so downloadable from their website (universaldependencies.org), alongside the others treebanks annotated according to the UD principles and set of relations.

RoRefTrees can be queried online using different tools, developed in different projects: at http://bionlp-www.utu.fi/dep_search, using SETS querying system, described at http://bionlp.utu.fi/searchexpressions-new.html; at http://lindat.mff.cuni.cz/services/pmltq/#!/home, using PML Tree Query, described at https://ufal.mff.cuni.cz/pmltq/doc/pmltq_doc.html; at http://clarino.uib.no/iness/page?page-id=iness-main-page, with the INESS (Rosén *et al.* 2012) infrastructure, described at http://clarino.uib.no/iness/page?page-id=inessdocumentation.

Figure 3. Searching RoRefTrees for verbs taking direct objects (left) and for verbs taking both direct and indirect objects (right)

We exemplify now searching for transitive verbs in RoRefTrees within INESS. The query phrase is: #x >dobj. The first page of the found verbs is displayed in Figure 3 (left) above, where one can see that this query has 5511 hits, that is there are 5511 occurrences of transitive verbs in the treebank. When considering their lemma, the number of such verb is 2621.The figure contains the most frequent forms, in reverse order of their frequency in RoRefTrees. In Figure 3 (right) we show the results of found when searching for verbs taking both a direct and an indirect object; the query is: (#x >dobj) & (#x >iobj).

A tree representation of one of the sentences containing such a verb is presented in Figure 4. The verb is the root of the sentence (*solicitat*), the dobj (*ani*) and the iobj (*i-*) are italicised in the figure.

The searching interface, the query language and the way trees are displayed differ from one project to the other. The user must get familiar with them and then working with the treebank will be a comfortable activity.

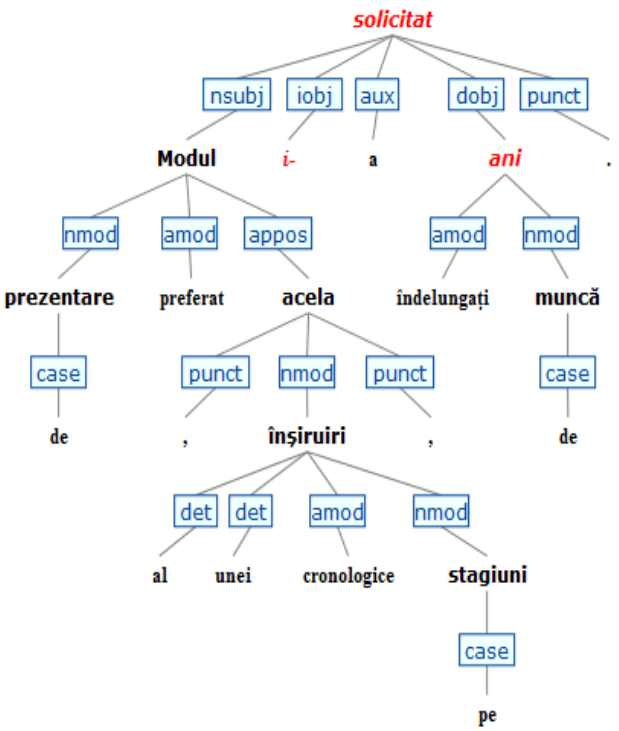

Figure 4. A tree representation of a sentence containing a verb with a dobj and an iobj relation (in red)

7. Conclusion

The work described here, of creating a Romanian treebank to be used for training and testing a Romanian parser, was done within an international

context: the UD project, version 1.4. Although quite new, it is rapidly evolving and the guidelines are now, at the time of this writing, at their 2.0 edition. The conversion of RoRefTrees from UD v 1.4 to the new annotation guidelines (v. 2.0) was made automatically, by the UD team. Each newly released version of the treebank is archived on the UD website.

This treebank for Romanian was created with an eye to diversity from various perspectives (style, domain, sentence type and sentence structure). Various linguistic studies can be based on this treebank and it can also be used for training and testing syntactic parsers for Romanian or language-independent ones.

References

Barbu Mititelu *et al.* 2016: Verginica Barbu Mititelu, Radu Ion, Radu Simionescu, Elana Irimia, Cenel-Augusto Perez, *The Romanian Treebank Annotated According to Universal Dependencies,* in *Proceedings of The Tenth International Conference on Natural Language Processing (HrTAL 2016).*

Erjavec 2012: Tomaž Erjavec, *MULTEXT-East: morpho-syntactic resources for Central and Eastern European languages.* In *Language Resources and Evaluation*, March 2012, Volume 46, Issue 1, p. 131–142.

Ion 2007: Radu Ion, *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD Thesis, Romanian Academy.

Irimia, Barbu Mititelu 2015: Elena Irimia, Verginica Barbu Mititelu, *Building a Romanian Dependency Treebank,* in: Corpus Linguistics conference (2015).

Nivre *et al.* 2007: Joakim Nivre, Johan Hall, Lens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov, Erwin Marsi, *Malt-parser: A language independent system for data-driven dependency parsing*, in "Natural Language Engineering", 2007, 13, p. 95–135.

Perez 2014: Cenel-Augusto Perez, *Linguistic Resources for Natural Language Processing.* PhD thesis, "Alexandru Ioan Cuza" University of Iasi.

Rosén *et al.* 2012: Victoria Rosén, Koenraad De Smedt, Paul Meurer, Helge Dyvik, *An open infrastructure for advanced treebanking.* In: Jan Hajič, Koenraad De Smedt, Marko Tadić, Antonio Branco (eds.), META-RESEARCH Workshop on Advanced Treebanking at LREC2012, p. 22–29.

Tufiş *et al.* 2008: Dan Tufiş, Radu Ion, Alexandru Ceauşu, Dan Ştefănescu, *RACAI's Linguistic Web Services.* In Nicoletta Calzolari *et al.* (Eds.) Proceedings of the 6th LREC, Marrakech, Morocco, European Language Resources Association (ELRA).

Modern Syntactic Analysis of Romanian

In this paper we present a Romanian treebank consisting of 9523 sentences. They were selected so that to cover more text styles, domains, sentence types and sentence structure types. They were automatically tokenized, part of speech tagged and lemmatised. The syntactic analysis of the sentences was done semi-automatically, following the principles and the set of relations from the Universal Dependencies project. This treebank is freely available for download and query from the Universal

Dependencies project site, where new versions of it, adapted to the new versions of the project guidelines, are also available. The treebank can be used both for linguistic investigations and for training and testing syntactic parsers for Romanian or language-independent ones.