International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 83–103
Received 24/01/2015, Accepted 27/02/2015, Final 20/12/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

A Romanian Dependency Treebank

CĂTĂLINA MĂRĂNDUC
Alexandru Ioan Cuza University of Iasi, Romania
Academic Institute of Linguistics, Bucharest, Romania

CENEL AUGUSTO PEREZ
Alexandru Ioan Cuza University of Iasi, Romania

ABSTRACT

The Romanian Treebank was created with manual and automatic manually checked annotation. The syntactic relationships were meticulously defined. We aim to affiliate our Treebank to Universal Dependencies, in this way some categories would become subclassifications. For the creation of this Treebank, we have built an annotation interface and a Romanian language dependent parser that works with statistical methods and whose accuracy is not satisfactory. This is why we intend to create another hybrid and rule-based parser. Its programming would include syntactic and semantic background for most verbs in the Romanian language. They will be extracted from Treebank and RoWN (lined up with PWN). For the missing verbs, we will introduce in the Treebank a big number of quotations from eDTLR (Romanian Thesaurus Dictionary). We will add to the Treebank a new layer with semantic annotation based on accurate criteria, starting with RoWN's, to which we will add interdictions. Another derived project is the creation of a multilingual aligned Treebank.

Keywords: Relationship, rules-based parser, semantic background, syntactic background, tree alignment, universal dependencies.

1. INTRODUCTION

Recent comparative studies have shown that besides English, all the other European languages have an insufficient degree of computerization. In terms of vocabulary, Romanian has an average level of computerization, by virtue of the existence of RoWN (Romanian WordNet) [3] aligned with PWN (Princeton WordNet), but it is ranked among the last when regarding the existence of annotated corpora approachable by researchers.

We intended to create a new Treebank for Romanian that would meet the urgent need of the computerization of Romanian language. The creation of the Treebank has started within a joint project shared by the Institute of Information Managementof the Romanian Academy and the Computer Science Faculty ofAl. I. Cuza University of Iasi. It started with a collection of 600 sentences annotated at the syntactical level. During the research in his PhD thesis, Augusto Perez [11] had extended it to 4,600 sentences in December 2014. This Treebank, called UAIC-RoDepTb, is balanced, containing complex sentences from all language registers. Some of Sentences are complex, with a varying extent, starting from 4 components to over 100.The total number of words and punctuation elements reached over 105000[1].

In these structures, using annotation conventions of the dependency grammar type, suited to the characteristics of the Romanian language, there were annotated a great number of relationships, whose names are similar to those of classical grammar. In addition to the verbal mandatory arguments, there were annotated a large number of optional circumstantial relations (called adjuncts or modifiers and generally no classified in the syntactic ontologies), which are useful for further semantic, pragmatic, discursive type of research.

[1] The data in this section were available for April 2015, when the article was communicated to CICLing conference. A recent assessment shows that the UAIC-RoDepTb reached 10,920 Sentences, and 200,764 words and punctuation elements.

2. CORPUS DESCRIPTION

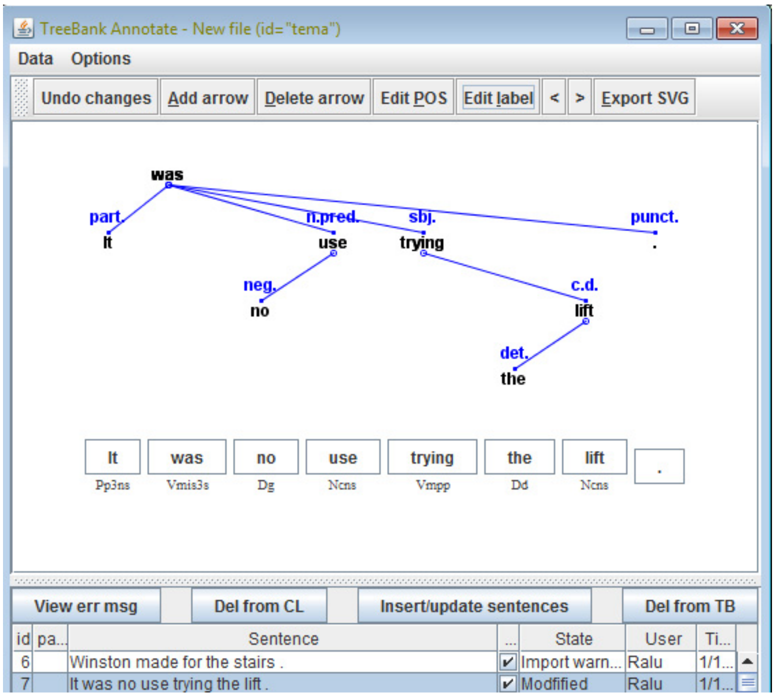
2.1. *Dimensions of corpus*

Unlike the other Treebankfor Romanian, made at the Faculty of Mathematics, University of Bucharest [8], and which had a comparably similar number of trees, our corpus has three times more syntactic units, which demonstrates that we have obtained a superior Treebank regarding the length and complexity of the syntactic structures.

Annotation was performed using an interface called the TreeAnnotator,[2] which allows viewing the arcs between node words or punctuation signs of the tree and inscribing the logos of syntactic relations as arcs labels. The graph obtained has each node positioned above the word from the subjacent initial array of signs;the morphological analysis label is visible as a result of automatic annotation of one of the two POS-taggers for Romanian language.[3]

[2] The interface is developed in the master degree paper by IustinDornescu.

[3] TTL POS-tagger from http://www.racai.ro/tools/text and UAIC POS-tagger from <http://nlptools.infoiasi.ro/Resources.jsp> are used from the morphologic previous annotation of sentences in the UAIC-RoDepTb.



id	pa...	Sentence	...	State	User	Ti...
6		Winston made for the stairs .	☑	Import warn...	Ralu	1/1...
7		It was no use trying the lift .	☑	Modified	Ralu	1/1...

Figure 1. *A graphviewedby the TreeAnnotator framework*

The annotation conventions used were determined by comparing the two language expert annotators' options and choosing the mutually agreed solutions in accordance with the language distinctiveness and the complexity of natural language phenomena that we were confronted with.

2.2. *Annotation conventions*

Hence an establishment was reached regarding a coherent system of punctuation annotation which, except commas marking the coordination, is subordinated to the head of the sub-tree that is isolated from the rest of the sentence by inverted commas, parentheses or commas. It may be an exogene structure, which comes from another emitter citation, which is grouped around a vocative without syntactic function, an optional dependency of a verb or an apposition. It can be seen that the punctuation annotation, imposed by the conventions of dependency Grammar theories, is justified as the punctuation signs have a defined syntactic-semantic role.

Another specific annotation convention regards the subordinate elements of elliptical regents. Our solution is different from that envisioned on the Universal Dependencies[13] site,and we consider it better. Example:

John won bronze, Mary silver, and Sandy gold. (1)

Their solution is to mark a relation (labeled *remnant*) in fact non-existent, between the three subjects on the one hand, and the three complements on the other hand (see Figure 2). In fact, they are part of different sentences and relations are established only between each subject and its object.

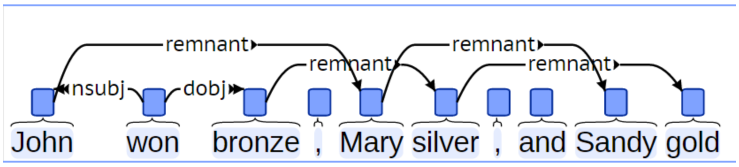


Figure 2. *Annotation of eliptics regents in English Treebank affiliated of Universal Dependencies*

We chose to use the label *remnant* for the relationship between the verb and the coordination elements, which in this sentence replace the elliptic regents, borrowing the properties of the root and becoming heads of pairs two and three of *subj* and *dobj* related (see Figure 3).

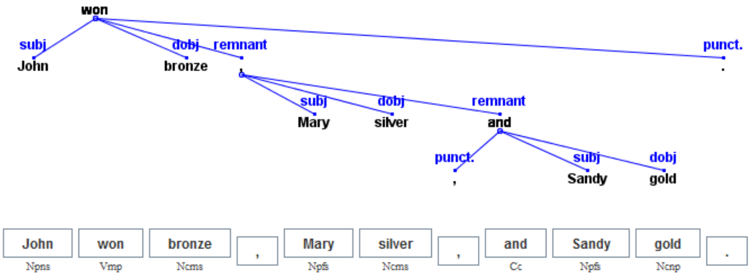


Figure 3. *Annotation of eliptics regents using Romanian Treebank conventions*

2.3. *Parsers*

For the automatic annotation, we used a FDG type parser, which was trained using the first manually annotated 1000 sentences, then with the bootstrapping method the automatically annotated sentences was manually corrected and the training was resumed with a gold corpus increased to 2000 and 3000 phrases. We then used another parser, similarly trained.[4] The evaluation ofparsers led to modest results (see Table 1 and 2). Both parsers developed variants of parser for Romanian built by J. Nivre [6-7].

In their article, Călăcean and Nivre [2] described asyntactic annotated corpus which resulted from the operation of their parser, which has 4,042 sentences with 36,150 tokens, punctuation excluded. The labeled attachment score for this parser had 88.6% accuracy, and the unlabeled attachment score had 92.0% accuracy for this corpus. The corpus includes short sentences with a simple structure, in a unique style, with political and administrative journalistic topics. Texts including complex ambiguities were avoided as much as possible and removed from their Treebank.

The evaluation of the two parsers on our Treebank led to the following modest results, although heuristic modules were used to improve the parser built by J. Nivre:

[4] FDG parser was developed in the graduation degree paper by Claudius Popa. The other parser was built by RaduSimionescu and can be found at http://nlptools.infoiasi.ro/WebFdgRo/.

Table 1. *FDG parser evaluation*

Metrics	Good trees	Label precision	Head precision	Both precision
Score	3,20%	55,95%	65,32%	61,03%

Table 2. *Evaluation of nlptools.uaic parser*

Metrics	Label precision	Head precision	Both precision
Score	62,02%	68,88%	58,56%

The poor results can be explained by the small size of the training corpus, the stylistic variety and the complexity of syntactic constructions. While other corpora have simple and typical examples specifically selected for an efficient training of the parser, the authors of the present Treebank, being both linguists, are concerned with the formulation of annotation conventions thorough and flexible enough to illustrate a wide variety of natural language and specific for Romanian linguistic facts.

The corpus illustrates the legal style, including texts of Acquiscommunitaire, Romanian and universal fictional texts, and also the journalistic style of Frame-Net translated into Romanian. The result of manual annotation can sometimes be the difference between relationships interpreted by the two annotators. We had not rigorously assessed the agreement between annotators using a statistical method. We preferred proceeding from mutual corrections and debate to establish a common view.

To get bigger corpus drive, would require a more effective automatic annotation, and to increase the accuracy of parser would require a bigger gold corpus for training. The two shortcomings (the reduced size of the corpus and the modest parser accuracy) are correlated. UAIC-RoDepTb has been developed too much through manual correcting of automatic parsing. In addition, there is an increasing interest among researchers, especially linguists, in the creation of processing tools for old Romanian, i.e., sixteenth and seventeenth centuries, Romanian' and this task would cause further difficulties for the training of a parser based on statistical methods. In the next stage of the research, we propose building a hybrid, statistic and rule-based parser, trained in distinct modules for contemporary and for old Romanian.

3. THE UNIVERSAL DEPENDENCY TREEBANKSFORMAT

3.1. *The need to unify the format of resources*

To make comparisons, to permit the reuse of our resources in various kinds of research, to get good results in our research, we aim at affiliating our Treebank to the Universal Dependencies group, founded in 2013, and to which were affiliated dependency Treebank corpora for 30 languages [13][5]. The categories of annotated relationships will be automatically translated into the categories proposed by this project, so some of our annotations will remain in a different layer of annotation. We will continue the process of annotation using the new labels for categories and subcategories.

To be implemented from a system of annotation to another, categories of the two systems should be in a relationship of equivalence or inclusion. There are cases in which the categories of our annotation system are in a relationship of intersection with Universal Dependencies ones.

We have carefully studied the annotation conventions in the English Treebank, which relationships were taken as examples on this site so as to compare them with those used by us. Each type has an intension, a definition of the relationship, and an extension, represented by the set of natural language facts that can be annotated with that relationship. It is important to study the relationship established between the set of relations defined by UD (Universal Stanford Dependencies) and the relations circumscribed by the UAIC-RoDep Tb conventions.

At a first glance, it would seem that the transposition will be easy, because there are many cases of equivalence between logos, or double inclusion between their extensions, for example: {appos}≡{ap.} {neg}≡{neg.} {parataxis}≡{incid.} {mark}≡{part.} {punct}≡{punct.} {vocative}≡{voc.}. But these are not the most important sections of the system. In other cases,

[5] http://universaldependencies.github.io/docs/#language.

there are differences between the sets of phenomena which are annotated. In other cases the sets are in inclusion or intersection relation, which is explained by the existence of theoretical and systemic differences that we try to synthesize and to comment.

3.2. *The distinct annotation of subordinate clauses*

The first observation is that, although clearly intended to establish a minimum number of syntactic categories, the same relationships are annotated differently when establishing a lexeme with his regent to where they are established between the same propositional and construction regent.

Examples:

*Cuvintele*lui au sens./His *words* make sense. (2)
Ceeacespune el are sens./ *What he said* make sense. (3)

In sentences (1, 2) we annotate the same relationship, sbj. on the line joining the underlined sequences of their regent. In contrast, according to the conventions of UD annotation, in sentence (2) the underlined word make *nsubj* relationship with the verb and in sentence (3), the relation of the underlined sequence is annotated with the *csubj* relationship (clause-subject). In our opinion, this is the same relationship with the same meaning and can be easily replaced with each other in any context. We notice that the label sbj. of our Treebank has an extension covers a lot of linguistic phenomena, and include the two sets referring to labels *csubj* and *nsubj* of UD annotation system (4). The same types of relationship are between the sets of expressions (5), due to the fact that our logo aux. is used in annotating to all types of auxiliaries.

$\{\text{nsubj}\}\subset\{\text{sbj.}\}$; $\{\text{csubj}\}\subset\{\text{sbj.}\}$ (4)
$\{\text{aux}\}\subset\{\text{aux.}\}$; $\{\text{auxpass}\}\subset\{\text{aux.}\}$ (5)

3.3. *Types of circumstantial modifiers*

As already mentioned, in our Treebank, which aims to further the basis for semantic annotation and discursive, argumentative and pragmatic research, great attention was given to establishing the types of Circumstantial Modifiers. They are annotated in UD system indiscriminately, without regard for their particular purposes and argumentative report that is set by the regent, placing emphasis instead on their morphological peculiarities, which however results from the previous automatic annotation with the POS-tagger. We do not know what theoretical arguments justify a return to the inferior morphologic level, since higher syntactic level should look, we believe, at more complex levels of the communicative organization including semantic, textual and discursive information.

Our system includes 14 categories: c.c.m. (modal circumstantial), c.c.t. (temporal circumstantial), c.c.l. (local circumstantial), c.c.cond. (conditional circumstantial), c.c.scop. (purpose circumstantial), c.c.cz.(cause circumstantial), c.c.cons. (consecutive or result circumstantial), c.c.conc. (concessive circumstantial), c.c.exc. (exception circumstantial), c.c.instr. (instrumental circumstantial), c.c.soc. (associative circumstantial), c.c.cumul. (cumulative circumstantial), c.c.opoz. (opposition circumstantial), c.c.rel. (relational circumstantial), categories of relationships that are interesting for further research and we do not intend to give up these distinctions once established and annotated; they will be kept in a different layer of annotation. To establish convergence with UD annotation system, as will be seen in what follows, circumstantial relations expressed by adverbs (i.e., adjuncts) are classified as *advmod*, while those expressed by nouns with a preposition to be annotated as *pmod* and those expressed by sentences are annotated with *clmod*. Consequently, a complicated system will result, with 14 intersections of each of the three types below: 14∙3=52intersectionsbetweencategories(6-8):

{advmod}∩{c.c.m.} ... {advmod}∩{c.c.rel.}; (6)
{advcl}∩{c.c.m.} ... {advcl}∩{c.c.rel.}; (7)
{pmod}∩{c.c.m.} ... {pmod}∩{c.c.rel.}. (8)

The system proposed by UD relations will use mandatory dependencies of really important verbs (called arguments), annotated with *nsubj*, *dobj*, *iobj*, which would be added *secobj* (secondary) *andagc* (Agent argument). Verbal circumstantial dependencies (called adjuncts) are considered optional, therefore less important, not reaching the basic structure of the sentence. But there are verbs for which certain circumstantial "adjuncts" are mandatory, which is another argument in favor of the syntactic-semantic significance of these categories. Examples:

A se deplasa de la Ana la Caiafa./
To move from Ana to Caiaphas. (9)

A dura de la oraunupână la oracinci./
To last for one hour at five. (10)

Camionulcântăreştepatru tone./
The lorry weighs four tons. (11)

In the Example (9), the verb *a se deplasa* "to move" involves two limits of space, so it has two mandatory dependencies c.c.l. In the Example (10), the verb a dura "totake" involves two time limits, so it has two mandatory dependencies c.c.t. In the Example (11), the verb a *cântări* "to weigh" has c.c.m. quantitative mandatory.

3.4. *Noun modifiers*

A similar situation emerges if we study the modifiers of the noun annotation in the two systems. The annotation conventions of UD can be: *amod*, *acl*, *nmod*, *pmod*, noting that the last category ascribes that dependence on verb. The focus is on annotation relationships based on their morphological realization, although, in our opinion, it is syntactically less important and in addition, marking this information is redundant, ranging in annotation POS tagger, preceding syntax.

In our annotation system, the noun dependents are not differentiated according to whether or not there exists a preposition, leading to junctions (13). The noun dependents are: a.adj. (adjectival attribute), a.subst. (noun attribute), a.pron. (pronominal attribute), a.adv. (adverbial attribute), a.vb. (verbal attribute). To make the transposition of noun modifiers from one format to another, again you have to take into consideration a number of relations of inclusion and intersection. In the system of UAIC-RoDepTb, *a.adj*. modifier includes numerals and determiners derived from pronouns, whereas *det*. (categories of determination) content only articles, whwn in the UD system *det* includes determiners derived from pronouns (14):

$\{\text{nummod}\}\subset\{\text{a.adj.}\}$ $\{\text{det}\}\cap\{\text{a.adj.}\}$ $\{\text{amod}\}\subset\{\text{a.adj.}\}$ $\{\text{a.adv.}\}$ $\subset\{\text{advmod}\}$; (12)
$\{\text{a.pron.}\}\cap\{\text{nmod}\}$ $\{\text{a.subst.}\}\cap\{\text{nmod}\}$ $\{\text{a.pron.}\}\cap\{\text{pmod}\}$ $\{\text{a.subst.}\}\cap\{\text{pmod}\}$. (13)

These situations can be resolved automatically just because the information is already morphologically annotated and it is likely that changes can be made without loss of information and without the need for another different layer of annotation, as in the case of complements.

3.5. *Annotation of prepositions and conjunctions*

UD uses annotation conventionsin which the conjunctions, prepositions, and marks of coordination are not considered head for the words that these connectors entered in the text. In our system, copulative verbs, prepositions and conjunctions are considered head, but this convention may be changed without great loss of information unless, except the case of examples similar with (1), in which the conjunctions or punctuation elements are instead of elliptical regents (translating by coordination their information).

But what seems highly inappropriate is the fact that the preposition annotation system is labeled UD case, because in some languages, (as French) a preposition forms the genitive case. It would be better to annotate it as mark. Prepositions introduce highly specialized semantic relations which are expressed not only by nouns or pronouns, but also by other parts of speech, like verbs, adverbs, for which the case category is not appropriate. In Romanian, the preposition does not expresses the case of nouns, but it requires, as a true regent, a specific noun case, and inflected forms are expressed by the enclitic definite article. There are illustrative examples:

El se ascundedupăcoteț./ He is hiding behind the cage. (14)
El se ascundeînapoiacotețului./ He is hiding behind the cage. (15)

Examples (14, 15) are synonyms. In (14), the preposition *după* "after" requires the accusative case and inarticulate form of noun. In (16), another preposition, *înapoia* "behind", requires the genitive case and articulate form of noun. The case is formed by the definite article *-ului* and not by the preposition. The label of this relationship should be, in our opinion, prep, or, if we decide to unify prepositions with subordinate conjunctions, *subord*.

Universal Conventions annotation should believe, to be the result of laborious consultations between computer scientists and linguists specialized in different natural languages, and not to unify categories of relations sacrificing semantic information in favor of the morphological one. It does not contain generalizations of some specific English features for all languages, such as those related to the expression of genitive case in French or the lack of reflexive in English. But since the unification of terminology and annotation formats is a high importance, we will submit the majority consensus and we adopt the system of relations of UD.

4. USING OUR TREEBANK FOR BUILDING NEW RESOURCES

4.1. *Inventory of predicate argument structures*

As we mentioned, we design the Treebank as a basis for more complex annotations. Outside to the standardization of the format and the increase of Treebank's size, we propose to create, through reuse of this corpus, some new resources for Romanian.

One of the development possibilities will be Treebank's enrichment through automatic annotation of semantic information extracted from RoWN.[6] This information will be assigned according to the lemma of the word, which was

[6] http://www.racai.ro/wnbrowser/

automatically annotated previously the syntactic annotation. Semantic annotations will be then corrected by experts.

After this preliminary stage, we use our Treebank to the development of a resource consisting of an inventory as comprehensively as possible of predicates arguments and adjuncts structure for Romanian. The predicates with their necessary or facultative dependencies will be extracted from the Treebank already annotated with syntactic and semantic labels. The corpus began to be created by computer scientists from RACAI (Romanian Academy Research Institute for Artificial Intelligence) by extracting such verbal patterns from RoWN. These types of syntactic-semantic structures [1] have been extracted for more than 500 verbs. Here's an example:

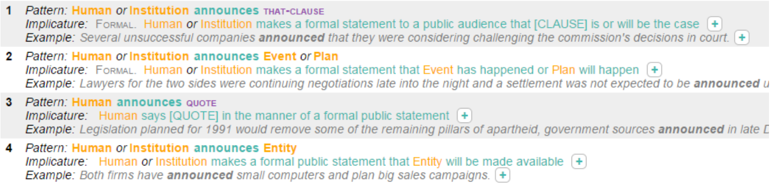
{mânca} nom*AG(person:1|animal:1)=acc*SUBSTANCE(food:1) (16)

The subject of the verb *amânca* "to eat" in (18), has the semantic role AG (ent), that can be satisfied by *animal* with meaning 1 in RoWN, by *person* with the meaning 1 in RoWN, or any noun in the nominative case (nom), which appears in RoWN as a hyponym of *person*: 1 or of *animal*: 1. The direct object in the accusative case (acc) has the semantic role SUBSTANCE and can be satisfied by the noun *hrană* "food" with the meaning 1 or by any of its hyponyms in RoWN. In parallel, at UAIC were built RoVerb-net, by the adoption of English Verb-net and by searching in Romanian corresponding examples for its categories of verbs [5]. The semantic roles of verbal group have been annotated by importing the Frame-net system of annotation in [15].

The new corpus, extracted by the Treebank, once it will be sufficiently representative for Romanian language verbs, will be used first in linguistic research, on the other hand in the natural language engineering, for programming a hybrid, statistic and rule-based syntactic parser. It is an absolutely necessary tool for increasing the size of the Treebank.

A long practice of automatic parsing error corrections has led us to the conclusion that the most affected by parsing errors are the structures containing misinterpretations in the upper place, at the root detection and his arguments required. Free word order in Romanian makes the parser to confuse the subject with the direct object or the predicate name. It is therefore necessary that the structural patterns of predicates contain rules in the form of prohibitions, i.e. syntactic-semantic dependencies that cannot be subordinated to that verb. The reflexive verb cannot have a direct object, for example. The confusion in the relations introduced by prepositions are more numerous in Romanian because there are more features and more occurrences of nouns preceded by the preposition than those with direct connection.

It is important to establish semantic categories as numerous as suitable for the purpose intended and as close to an international standard. In addition to the syntactic categories found in verbal-semantic structures extracted from RoWN, we can use PDEV[7] (Dictionary Pattern of English Verbs) that provides semantic frames for 5,602 verbs. The shape can be seen in Fig. 4. These structures protocols that were used to create the English Verb-net, and also the classification made in [10] can be largely translated into Romanian or can be taken as a model for creating patterns whose structure is not similar in both languages, Romanian and English.



1 *Pattern:* **Human** *or* **Institution announces** THAT-CLAUSE
Implicature: FORMAL. Human *or* Institution makes a formal statement to a public audience that [CLAUSE] is or will be the case +
Example: *Several unsuccessful companies* ***announced*** *that they were considering challenging the commission's decisions in court.* +

2 *Pattern:* **Human** *or* **Institution announces Event** *or* **Plan**
Implicature: FORMAL. Human *or* Institution makes a formal statement that Event has happened or Plan will happen +
Example: *Lawyers for the two sides were continuing negotiations late into the night and a settlement was not expected to be* ***announced*** *u*

3 *Pattern:* **Human announces** QUOTE
Implicature: Human says [QUOTE] in the manner of a formal public statement +
Example: *Legislation planned for 1991 would remove some of the remaining pillars of apartheid, government sources* ***announced*** *in late D*

4 *Pattern:* **Human** *or* **Institution announces Entity**
Implicature: Human *or* Institution makes a formal public statement that Entity will be made available +
Example: *Both firms have* ***announced*** *small computers and plan big sales campaigns.* +

Figure 4. *Syntactic semantic structure of the verb announce in the PDEV*

For the Romanian language, the structure of Figure 4, which add syntactic information will look like this:

[7] http://www.pdev.org.uk/#browse?q=;f=C

{anuța:1} *nsubj* [HUMAN or INSTITUTION] announces *cs* [CĂ] *dobjcl* [CLAUSE]. Example: (17)
Municipalitateaanunţăcăimpozitelevorcreşte./ The Municipality announce that taxes will increase.
{anuța:2} *nsubj* [HUMAN or INSTITUTION] announces *dobj* [EVENT or PLAN] *iobj* [HUMAN 2].
Example: Emil anunţăcăsătoriasaprietenilor. / Emile announces his marriage to friends.
{anuța:3} *nsubj* [HUMAN] announces *dobj* [DATE]
Example: Meteorologulanunţă 32° C. / The weatherman announces 32° C.

The fourth situation in Figure 4 has no direct parallel in Romanian. But as dependency grammar conventions do not provide transformations, we have to include the appropriate distinct patterns as the examples above (17), with the same verb in the passive, reflexive, impersonal voice, if they are present in Romanian. Here's one of them:

{anunța:4} *nsubjpass* [EVENT or PLAN] is announced *prep* [DE (CĂTRE)'by'] *cag* [HUMAN/INSTITUTION]
Example: Căsătoriaesteanunțată de (către) Emil./ The marriage is announced by Emile. (18)

In a recent article describing a similar resource for Italian [9], the authors show that such syntactic semantic patterns have been extracted by lexicographical methods from large text corpora. So, they detect new possible situations that did not fit the pattern set, such as:

Pattern: [HUMAN1] annuncia [EVENT] a [HUMAN2]
Corpus lines: L'altoparlanteannunciaval'arrivodeltreno/ The speaker announced the arrival of the train. (19)

The authors consider that (19) is a type mismatch, wrong framed as an example to existing types. It would be necessary to introduce a new pattern:

[DISPOSITIF, MACHINE] annuncia [EVENT] a [HUMAN] (20)

Besides the Treebank corpus from which we can extract sentences containing the verb for which we have to build anew pattern structure, we have another resource for Romanian, eDTLR (Electronic Romanian Language Thesaurus Dictionary)[8] obtained in a project in which more Romanian academic institutions participated, by parsing the 32 printed books, scanned, converted into text by OCR-ization and corrected by experts. From another corpus related to the same project, the bibliographical sources of eDTLR, we have already randomly selected 1000 sentences that were introduced in the Treebank and which have the advantage of a balanced selection of the most representative Romanian texts of every style.

The dictionary contains over 2 million quotations arranged chronologically and by the numbers of meanings of every word-entry, following the definition. We have to select from these quotations those related to verbs that do not have syntactic-semantic structures in the RoWN, they will be introduced in the Treebank corpus enriched with semantic annotations. Unlike the examples (17-20), the examples of the new resource will have a tree form, with exact indication of head for mandatory and optional dependencies.

Type prohibition rules attached to patterns of verbs for constructing rules-based syntactic parser are selected from another corpus, unfortunately also rising, i.e. the corpus of automatic syntactic parsing errors corrected by experts. The rules will be obtained by generalization of the most common errors, and they will have the form:

[NOT] *refl* [SE] announces *dobj* [EVENT or DATE]. (21)
Se anunță o vremefrumoasă. /It announces nice weather.

In the Example (21) in Romanian it is about not a *refl* but an *impers* value and the syntactic function required is the *nsubj* or

[8] Built between 2007 – 2010, in a project financed by Romanian Government and coordinated by UAIC-FII (https://consilr.info.uaic.ro/edtlr/wiki/index.php?title=Digitalizing_the_Thesaurus_Dictionary_of_the_Romanian_Language).

csubj achieved through a sentence, which will be established in other patterns that will result from the corpus of quotations introduced in the Treebank.

4.2. *Aligned corpus 1984.en.ro.fr*

Since our Treebank already contains nearly 1,000 Sentences coming from Romanian version of the project alignment Translations 1984[9] we decided to build a corpus which contained the annotated sentences in the Romanian version of 1984 and the parallel sentences in English and French, annotated with respect of our Treebank conventions. The corpus, began in December 2014 by Master of computational linguistics students of Faculty of Computer Science of Al. I. Cuza University of Iasi, has 250 Sentences in each of the three languages, each containing about 6,500 tokens. Corpus size could be increased by aligning other French and English Sentences from the Romanian already annotated or by adding other languages.

In Figure 1 there is a tree belonging from the aligned English corpus that was annotated too in the Tree Annotator interface used to create the Romanian Dependency Treebank. Fig. 5 shows the tree-structure of sentences aligned with it.

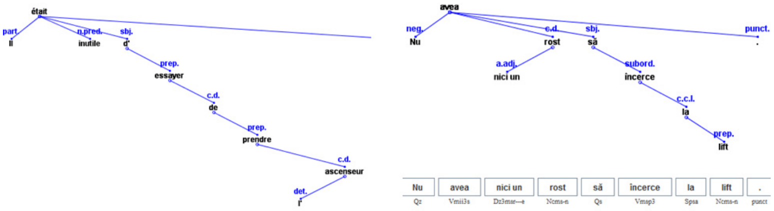


Figure 5. *Trees alined with the tree from Figure. 1.*

This project is important for the EBMTs (Example based machine translations) by introducing tree structures for the language source and for the language target into the memory of

[9] MTE (MultiText-East, morpho-syntactic manually annotated) (nl.ijs.si).

translation software, or to evaluate the quality of automatic translations, or to extract rules for rule-based machine translation. It is required already a program that aligns syntactic relations structures of the trees in different languages and automatically displays the differences found, like the program described in [14].

The study of aligned trees shows us structural syntactic differences between languages and then we can make assumptions on the most appropriate format annotation that would have to use a universal system of categories as the UD. At the same time, we see which format is readily convertible to other format that should have a system of dependency Treebanks. Although permanent concern us the most appropriate formalism to the Romanian languages pecific structures, we should avoid as far as possible the use of convertible difficult relationship in the annotation system conventions of other languages.

5. CONCLUSIONS

In this paper we tried to demonstrate that it is extremely important, especially for a small movement language with scientific research underfunded, but not limited to such a language, to reuse existing resources for building new resources, with a higher degree of complexity of the annotations. For this, we need most often towards compatibility annotation formats new or old. It is necessary to create tools to make unification of annotations, conversions, or simplifications, tools organized in chains of automatic processing. At our faculty such operations are carried out through the hierarchical meta-system of tools and resources ALPE (Automated Linguistic Processing Environment) [4], [12].[10]

In this context, the task of linguists, besides correcting the errors of natural language processing programs, is precisely to carry out studies concerning not only the logos of private correspondence with other standardized universal logos but also about the transposition of resources in a format based on a

[10] http://creativecommons.org/licenses/by-nc-sa/3.0/

particular set of annotation conventions into another format based on other annotation conventions.

REFERENCES

1. Barbu-Mititelu, V. 2013. *Derivational Semantic Network for Romanian*, Bucharest, National Museum of Romanian Literature Press.
2. Călăcean, M. & Nivre, J. 2008. Data-driven Dependency Parsing for Romanian, Acta Universitatis Upsaliensis (pp. 65-76).
3. Cristea, D., Mihăilă, C., Forăscu, C., Trandabăţ, D., Husarciuc, M., Haja, G. & Postolache, O. 2004. Mapping Princeton WordNet Synsets onto Romanian WordNet Synsets. In *Romanian Journal on Information Science and Technology*, *Special Issue on BalkaNet*, 7, 125-145.
4. Cristea, D. & Pistol, I. C. 2008. Managing language resources and tools using a hierarchy of annotation schemas. In proceedings of *the Workshop on Sustainability o f Language Resources, LREC, Marakesh*. (2008)
5. Curteanu, N., Moruz, M., Trandabat, D., Bolea, C. & Dornescu, I. 2006. The structure and parsing of Romanian verbal group and predicate. In proceedings of the *4th European Conference on Intelligent Systems and Technologies, ECIT.*
6. Hall, J., Nivre, J. & Nilsson, J. 2006. Discriminative classifiers for deterministic dependency parsing. In proceedings of the *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, Main Conference Poster Sessions.
7. Hall, J. & Nilsson, J. 2007. CoNLL-X Shared Task: Multi-lingual Dependency Parsing, MSI report 06060. Växjö University, School of Mathematics and Systems Engineering. (2007)
8. Hristea, F. & Popescu, M. 2003. A dependency grammar approach to syntactic analysis with special reference to Romanian. In F. Hristea & M. Popescu (Eds.), *Building Awareness in Language Technology* (pp. 9-34). Bucharest: University of Bucharest Press.
9. Jezek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. 2014. Structures for linguistic analysis and semantic processing. In proceedings of *LREC* (pp. 890-895).
10. Levin, B. 1993. *English Verb Class and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

11. Perez, A.-C. 2014. Linguistic Resources for Natural Language Processing. PhD thesis, Al. I. Cuza University, Iași.
12. Pistol, I. C. & Cristea, D. 2009. Managing metadata variability. *Milan: Proceedings of the 6th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2009* (pp. 111-116).
13. Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., & Žabokrtský, Z. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In proceedings of *LREC*. (2014)
14. Sanguinetti, M., Bosco, C. & Cupi, L. 2014. Exploiting catenae in a parallel Treebank alignment. In proceedings LREC 2014, 1824-1831.
15. Trandabăț, D. 2010. Natural Language Processing Using Semantic Frames, PHD Thesis, Computer Science Faculty, Alexandru Ioan Cuza, University of Iasi, Romania. Available online: <http://students.info.uaic.ro/~dtrandabat/thesis.pdf>.

ACKNOWLEDGEMENTS

Our gratitude goes to computer scientists in natural language processing specialists that we work with and without which our linguistic research would not be possible or would take place today as in the past century, during the decades by listing books in libraries and with more modest results than those now we can get in a few months.

CĂTĂLINA MĂRĂNDUC
FACULTY OF COMPUTER SCIENCE,
ALEXANDRU IOAN CUZA UNIVERSITY OF IASI, ROMANIA.
ACADEMIC INSTITUTE OF LINGUISTICS, BUCHAREST, ROMANIA.
E-MAIL: <CATALINA.MARANDUC@INFO.UAIC.RO>

CENEL AUGUSTO PEREZ
FACULTY OF COMPUTER SCIENCE,
ALEXANDRU IOAN CUZA UNIVERSITY OF IASI, ROMANIA.
E-MAIL: <AUGUSTO.PEREZ@INFO.UAIC.RO>