A Multiform Balanced Dependency Treebank for Romanian

Mihaela Colhon
University of Craiova
Department of Informatics
Craiova, Romania
mcolhon@gmail.com

Cătălin Mărănduc
“Al. I. Cuza” University,
Iași, Romania
Institute of Linguistics
“Iorgu Iordan Al. Rosetti”
Bucharest, Romania
catalinamaranduc@gmail.com

Cătălin Mititelu
Bucharest
Romania
catalinmititelu@yahoo.com

Abstract

The UAIC-RoDia-DepTb is a balanced treebank, containing texts in non-standard language: 2,575 chats sentences, old Romanian texts (a Gospel printed in 1648, a codex of laws printed in 1818, a novel written in 1910), regional popular poetry, legal texts, Romanian and foreign fiction, quotations. The proportions are comparable; each of these types of texts is represented by subsets of at least 1,000 phrases, so that the parser can be trained on their peculiarities. The annotation of the treebank started in 2007, and it has classical tags, such as those in school grammar, with the intention of using the resource for didactic purposes. The classification of circumstantial modifiers is rich in semantic information. We present in this paper the development in progress of this resource which has been automatically annotated and entirely manually corrected. We try to add new texts, and to make it available in more formats, by keeping all the morphological and syntactic information annotated, and adding logical-semantic information. We will describe here two conversions, from the classic syntactic format into Universal Dependencies format and into a logical-semantic layer, which will be shortly presented.

1 Introduction

The annotation of UAIC-RoDia DepTb1 was started in 2007, prompted by the need to complete the lack of corpora for Romanian with a high degree of annotated data. The creator, Augusto Perez, 2014, used classical tags with the intention to use the treebank for teaching purposes, in preparing students for exams.

Since then, there have created more resources and processing tools. Their creators are all interested in standard contemporary language, overlooking the complex structures, the originality of expression. Their aim is to obtain superior accuracy by processing simple texts. A big corpus for contemporary standard language (CoRoLA) has been created (Tușiș et al., 1998). It includes publications obtained from editors, spoken language and also a treebank consisting of 9,500 sentences, affiliated with the UD (Universal Dependencies) group. UAIC-RoDepTb is a balanced corpus, having more styles. We contribute our 4,500 sentences, which are in contemporary standard language, to the Romanian Treebank affiliated with UD.

However, the UD group includes all kinds of Treebanks, balanced, or containing old languages, Social Media, and so on. The initial purpose of this group was to build a universal parser. The common features of all the languages have been highlighted, admitting the peculiarities only as sub-classifications that can be taken into consideration or not, according to the user’s wish. More and more treebanks for over 30 languages have been affiliated. The uniformity of the flexible annotation format has created the possibility of multiple alignments, useful in Machine Translation and comparative language studies.

UAIC-RoDia DepTb can participate to this project as another treebank, for non-standard Romanian. Our Treebank has now 19,825 sentences in progress, in the UAIC classic syntactic convention of annotation. If we subtract the 4,500 sentences already introduced into the UD, there remain 15,325 sentences to be transposed in UD conventions: 2,575 sentences from chat, 1,230

---

1UAIC-RoDia = ISLRN 156-635-615-024-0
from regional folk poetry, 6,882 in Old Romanian, a Gospel published in 1648 and a codex of laws issued in 1818, Romanian fiction, a novel by Matthew Caragiale *The Princes of Old Palace*, written in 1910, and 3,894 quotations from the thesaurus dictionary or its bibliography, containing poetry by known authors, too. The section already introduced in UD contains fragments from the Romanian version of Orwell’s *1984* novel, sentences from the Aquis Communitaire laws, from Wikipedia in Romanian, and from the Frame Net (Baker et al., 1998), translated in Romanian. The sentences with complex structure have not been avoided, neither have short, elliptical ones. The average frequency is high, some sub-corpora have 25-29 words per sentence, and the general average frequency is 19.91 words per sentence.

The differences between the UAIC annotation conventions and UD ones have been shown in previous work (Mărarăncu and Perez, 2015), (Mititelu et al., 2015). Shortly, it’s about the annotation of functional words, considered in UAIC as heads, such as the copulative verbs, while in UD they are subordinated, and another convention for the annotation of coordination. Also, UAIC treebank has the same tag to mark a relation expressed by a word or by a subordinate clause, and UD use tags as: `ccomp`, `csubj`, `advc`, where the first `c` or the last `cl` means clause.

The annotation in the UAIC conventions was not renounced, for reasons of continuity. There has also been created an Old-Ro-POS-tagger for the complete morphological analysis of Old Romanian (starting with the sixteenth century) (Mărarăncu et al., 2017b). This classic, annotated with rich information format can be considered as the pivot from which the other two formats have to be automatically obtained.

The logical-semantic format that is actually used to capitalize the semantic information present in syntactic and morphological UAIC-RoDepTb annotations is described in Mărarăncu et al. (2017a). It has similarities with the Tectogrammatic layer of the Prague Dependency Treebank (PDT) (Bohmová et al., 2003) and Abstract Meaning Representation (AMR) (Bănescu et al., 2013).

In this paper, we briefly describe the semantic format, showing how most classic-syntactic tags can be automatically transformed into semantic ones. Some of the syntactic tags are highly ambiguous from the semantic point of view, and they are manually transformed using an interface that has drop-down lists (Hociung, 2016).

We will describe a program called TREEOPS that automatically transforms non-ambiguous syntactic tags. We shall give examples of rules for the transformation of syntactic into semantic tags, and then, we shall give examples of rules written in the same program, for transforming the UAIC syntactic structure of trees into the UD one.

2 Related Work

2.1 Dependency Treebanks

The UAIC-RoDia DepTb is annotated in Dependency Grammar, a flexible formalism founded by Tesnière (1959), Kern (1883) and actualized as Functional Dependency Grammar by Tapanainen and Jarvinen (1998), Mel’čuk (1988). Actually, a big number of corpora in the world have adopted the same formalism. All these corpora exist only if the work is going on; if they grow organically, if they have the flexibility to change their format into another, adopted by more universal resources, and the creators always try to introduce more refined annotations of linguistic phenomena. Otherwise, the amount of annotated information or of texts introduced is exceeded by other resources, and the format becomes obsolete, difficult to compare with the new modern ones. Consequently, the users forget this resource and prefer another.

In 2003, the PDT authors described the three level structure of their treebank and the Tectogrammatic level (that includes semantic, logical and syntactic information) (Bohmová et al., 2003). The PDT authors publish their updates every two years (Bejček et al., 2013). They have for a long time been interested in semantics and its links with syntax (Šgall et al., 1986).

BulTreeBank is another big member of the UD community. This treebank has been previously annotated in the HPSG grammar and automatically transposed into Dependency formalism. The authors are also interested in semantics (Simov and Osenova, 2011).

For the PENN Treebank, the third version was available in 1999 (Marcus et al., 1999). Actually, this treebank is also involved in semantics research or in the annotation of entities and events (Song et al., 2015).
2.2 Semantic Annotations

In the UAIC NLP group, about 1,000 sentences from the English FrameNet have been translated in Romanian (Trandabăț, 2010). Retaining the semantic annotation from the English FrameNet, the author has made a first set of semantic annotations on Romanian sentences. Just as the English FrameNet (Baker et al., 1998), these annotations only cover the core structure of the sentence, called Semantic Frame.

The similarities between the system of semantic annotations presented here and the AMR (Abstract Meaning Representation) logical categories (Bănărescu et al., 2013) are obvious. However, there are also important differences, since the resulting graph of the AMR semantic annotation is not a dependency tree, and the nodes are not words, but concepts.

The similitudes with the Tectogrammatic layer of the PDT (Prague Dependency Treebank) are more important. They also have, like in our semantic system, categories for annotating the semantic information existing in the exclamatory, or interrogative form of the sentence, in the modality and the tense of verbs, considering the punctuation or morphological annotation as containing semantic information.

3 The UAIC-RoDepTb Logical-Semantic Format

3.1 Short Characterization of the UAIC Semantic Format

The circumstantial information gives us indications about the state of communication: succession in time, past, future, space: TEMP, PAST, FTR, LOC, or information regarding logical relationships: cause, purpose, consequence, opposition, concession, condition, exception, cumulation, association, reference, restriction, result: CAUS, PURP, CSQ, OPPOS, CNCS, COND, EXCP, CUMUL, ASSOC, REFR, RESTR, RSLT, respectively.

Information on the names of objects or actors is derived from the classification of the pronouns: Appurtenance, Deictic, Emphatic, Undefined: APP, DX, IDENT, UNDEF. The classification of the functional words (considered as connectors) give us a classification of relations. There are six types of connectors: CNADVS, CNCNCL, CNCONJ, CNDISJ, CNSBRD, CNCOP, the first four being coordinating connectors, the fifth being subordinating connector, and the last being the copulative verb “to be”, considered as connector between the subject and their description rendered by the predicative noun. As a logical point of view, the different type of connectors mark relations of inclusion, reunion, disjunction, intersection, particularization or generalization of properties, between the set of things which the linguistic signs refer to. The truth value of sentences par rapport to the real World can also be established, and it receives modal values such as: existence, possibility and necessity.

There is also information regarding the reality of the action: optative, uncertain, potentiality, generic, dynamic: OPTV, UNCTN, POTN, GNR, DYN. Five quantifiers are considered: necessity, possibility, existence, universality, and negation: QNECES, QPOSSIB, QEXIST, QUNIV, QNEG. Other information qualifies the type of the communication: addressee, blame, greeting, politeness, interrogative, exclamatory, incidence: ADDR, BLAM, GREET, POLIT, INTROG, EXCL, INCID, respectively. The last tag marks the change of the emitter and receiver roles, i.e. it marks another communication state; the first and the second persons refer to other characters in the INCID text, isolated by NOAPP punctuation, than in the rest of the sentence. This is very important for a future program that would automatically establish the co-references. Also this information about the communication state is important for future pragmatics research.

EQUIVHEAD is the mark for the ellipsis. This mark allows a connector or punctuation element to represent a copy of the meaning of its head and to have dependents with a similar, and symmetrical structure. The expletive is classified according to the value it repeats: EXPL:BEN, EXPL:EXPR, EXPL:OBJ, EXLP:DFND, EXPL:PAT, EXPL:RCPR, EXPL:APP, EXPL:EXP. This annotation is also in view of a future program that would automatically establish the co-references.

The punctuation elements also have a semantic value in our system. The semantic value of punctuation elements is shown in Druguş (2015). The final ones are marks for the form: INTROG, EXCL, END. The non-final are annotated as CNCONJ if they mark the coordination, or NOAPP (non-appurtenance) if they are subordinated to an INCID, ADDR or interjection (ALERT, IMIT,
AFF) head. The others are markers of dislocation (topical changes, frequent in Old Romanian) or elaboration (structures providing additional information that can be omitted): DISL, ELAB.

Finally, since not all sentences contain events, we have made a classification of the types of scenarios that govern the roles in the sentence (see Table 1 3.1).

### 3.2 Syntactic Relations Without Semantic Ambiguity

The classical syntactic format of the UAIC-RoDepTb has 14 kinds of circumstantial modifiers, carefully supervised by linguists. Actually, the time and space automatic annotation is at the core of the computational linguists interest; our corpus can be a useful training corpus for future automatic parsing of the communication situation circumstances. The second edition of the workshop on “Corpus-based Research in the Humanities” (CRH) Viena, January 25th-26th 2018, will have a special topic concerning time and space annotation in textual data.² The annotation of the time and space is semi-automatically rendered; the circumstantial modifiers are automatically annotated, but words with the same meaning can have syntactic ambiguous values, as nominal modifier or prepositional object, and must be manually annotated.

Besides the circumstantials, other syntactic relations that can be transposed in a unique semantic tag are: superl., “superlative”, “=SUPER”, comp., “comparative”, “=COMP”, ap., “apposition”, =RSMP, incid., “incident”, =INCID, neg., “negation”, QNEG, voc., “vocative”, =ADDR,

| Judgment | nsubj | dobj | npred | other |
|----------|-------|------|-------|-------|
| Process  | ACT   | RLST | -     | -     |
| Performance | PERFR | PERF| QLF  | -     |
| Actantial | ACT   | PAT  | -     | BEN   |
| Experience | EXPR  | EXP  | -     | BEN   |
| Comunic. | EMT   | CTNT | -     | RCPT  |
| Definition | DFND  | DFNS | CNCO |CNCO |
| Chang.idnt | DFND  | DFNS | CNCO |CNCO |
| Characteriz | CTNT  | QLF  | CNCO |CNCO |
| Existence | QEXIST| -    | LOC.TME | |

Table 1: The semantic roles depending on the type of judgment.

c.ag., “agent complement”, =ACT. Our classical syntactic tagset has 44 relations; having 14 circumstantial modifiers and the 7 relations above, it results that almost half of them (21) are semantically monovalent. This is a consequence of the fact that the classic format contains a lot of semantic information.

### 3.3 Syntactic Relations with Semantic Ambiguity

Besides the relationships with a single semantic interpretation, it is also possible to make automatic transformations by rules with more conditions, using morphological information or word lemma. Examples:

- The syntactic tag det. can be transposed in the semantic tag DX “Deixis”, if the morphological analysis (POS-tag) begins with Td “Demonstrative article”, or in the semantic tag UNDEF, if the POS-tag begins with Ti “undefined article”, or in the semantic tag DEF if the morphological analysis begins with Tf “definite article”, or in the semantic tag APP “appurtenance” if the morphological analysis begins with Ts “possessive article”.

- The syntactic tag aux. “auxiliary” can be transposed in the following semantic tags: OPTV or PAST, if the form of the auxiliary (and of the conjugate verb) indicate(s) the verbal conditional or the past tense, FTR, if the auxiliary is vrea “will” or PASS “passive”.

For similar situations, the tags can be automatically transposed from the syntactic format into the semantic one by writing rules with two or three conditions in the TREEOPS program (see below).

### 3.4 Syntactic Tags Semantically Polyvalent

Syntactic relations classified by morphological criteria, such as: a.subst. (noun attribute), a.vb. (verbal attribute), c.prep. (prepositional complement), have a great semantic ambiguity, and so do those in the Table 1, the column 2 and 3. They can have almost any semantic value. We automatically modify the monovalent relations, for which we write rules with one condition. Other relationships that depend on several factors need rules with more conditions, and the structure of the trees, too. TREEOPS makes all the changes.
which are not ambiguous, in accordance with the rules introduced, and the ambiguous relations remain unchanged (as syntactic tags). At this moment, they are manually changed by an expert, simultaneously with supervising automatic annotations. In this way, we are building the training corpus for a semantic parser, see (Aho and Ullman, 1972). The training corpus has now 5,025 sentences in Old Romanian and 1,130 sentences in Contemporary Romanian; we must increase and balance it, adding approximately 3,500 sentences in Contemporary Romanian.

We will use a statistical semantic parser. It will receive the documents having all the possible automatically processed transformations (consequently having mixed syntactic and semantic relations), and will statistically transform the syntactic ambiguous ones into semantic relations. It will be similar to a mixed parser, first the rules written in the TREEOPS program will be applied, to make the transformations non-ambiguous, and then the parser will receive the output of TREEOPS program (that are correct since the resulting from the supervised syntactic layer) so as to solve statistically the ambiguous syntactic relations.

The supervision of the automatic transformation simultaneously with the manual choice of semantic values for the ambiguous syntactic relations is performed currently by using an interface called Treebank Annotator (Hociung, 2016) that has drop-down lists for each feature in each of the three formats: UAIC, UD and semantic. The format is selected from a list of options before opening a document.

4 The TREEOPS Program. Rules for Transformations

4.1 Presentation of the Program

The automatic transformations are done using a tool called TREEOPS. It is an all purpose rule-based XML transformer, i.e. it is able to produce a new XML structure, having an XML as input and using a customized set of rules. It defines a new simple language for XML transformations, where each rule is described by the following pattern:

\[ \text{selector} \rightarrow \text{action} \]

During a transforming process the XML is traversed node by node and the TREEOPS rule is converted into an \textit{if-then} statement as follows:

\begin{verbatim}
if (selector matches node) then action
\end{verbatim}

TREEOPS requires the \textit{selector} to be an XML Path Language\(^3\) (XPath) expression and the \textit{action} an internal defined action that can take parameters, as for example, \texttt{changeAttrValue(<new_value>)} which changes the value of the current XML attribute. In fact, TREEOPS uses the power of the XSLT language by transforming the set of rules into an XSLT template that will be applied to the input file to obtain the new desired structure.

For example, the rule defined as:

\begin{verbatim}
/word[@deprel='superl.']/deprel =>
changeAttrValue('SUPER')
\end{verbatim}

becomes an XSLT template:

\begin{verbatim}
<xsl:template match="/word[@deprel='superl.']/deprel">
  <xsl:call-template name="changeAttrValue">
    <xsl:with-param name="new_value" select="SUPER"/>
  </xsl:call-template>
</xsl:template>
\end{verbatim}

where the \texttt{changeAttrValue} template is pre-defined as:

\begin{verbatim}
<xsl:template name="changeAttrValue">
  <xsl:param name="new_value"/>
  <xsl:attribute name="{name(.)}"
    select="$new_value"/>
</xsl:attribute>
</xsl:template>
\end{verbatim}

4.2 Rules for the Transposition of Classical Syntactic Format into Semantic One

For the 21 tags with a unique semantic value, we have written 21 rules with a unique condition. Example: the rule below transforms the negation into a logical quantifier:

\begin{verbatim}
/word[@deprel='neg.']/deprel =>
changeAttrValue('QNEG')
\end{verbatim}

The rule exemplified below has two conditions. It changes the syntactic value coord. in a connector for the reunion. We classified the coordination relations in four logical categories, taking into account the lemma of the coordinating conjunction: reunion (CNCONJ), adversative (QNADVS) (opposition where the related entities do not exclude

\(^3\)https://www.w3.org/TR/xpath/
each other), disjunction (QNDISJ i.e. exclusion, and conclusive. The other three are written in the same way.

To give an example for a rule with three conditions, the rule of the PAST relationship attribution depends on the aux. syntactic relation, the word form of the auxiliary, and the postag of the main verb form, being a compound time. The information needed has been previously annotated in the syntactic format.

4.3 Rules for the Transposition of Classical Syntactic Format into UD Format

To transpose the classic syntactic format into UD conventions, we have formulated another set of rules in a similar way. Here’s an example of a rule with one condition:

As a general observation, in the first case, we need to transpose a syntactic tag set of 44 classes into a semantic tag set of 95, so more conditions have to be done and more cases remain to be manually solved. On the other hand, the transposition of the UAIC syntactic tag set of 44 tags into the UD tagset of 28 common tags is a simpler operation, based on unifications. Of course, these transformations will also need to be supervised.

5 Structural Transformations of Trees

We decided that the structural transformations should be applied both to obtain the UD format and also to the semantic format, to make it more accessible to alignments or comparisons with other treebanks. The subordination of prepositions was a structural change also applied to the PDT Tectogrammatic layer. Establishing relationships between meaningful (also called self-semantic) words is more appropriate for the semantic analysis of the sentence.

Figure 1: “To the market, to the pharmacy” before and after the subordination of prepositions.

5.1 Subordination of the Relational Words to the Word Which It Introduces

The subordination of prepositions to the word which they are introducing is the simplest and most frequent of the operations. The order of the operations is not random. Being the most common, usually located at the tree leaves, it must be the first transformation.

The rule for the first transformation is described in pseudocode in listing 1 and the result can be seen in Figure 1.

Listing 1: Transformation 1

| If word1[@id="x", @postag="Sp*"] and word2[@head="x"] |
| then |
| word1/@head ← word2/@id |
| word2/@head ← word1/@head |

The subordination of relational words which introduce the subordinated sentence is similar to the first; the aim is to subordinae the relational words in a complex sentence. This rule must be applied after the first one, because it is no longer applied to the leaves, but to the upper structure of the tree.

The rule is almost the same: if there is a word1 with the id="x" and the postag="Cs*", “Pw*”, “Rw*”, “Qs” and a word2 with head="x" (usually the next word), then change the head of word1 with the id of word2 and the head of word2 with the head of word1.

If the subordinate word is a relative pronoun preceded by a preposition, then the preposition has already been subordinated to it by the first trans-
formation. Consequently, we must introduce a restriction of the type with the head="x" and without the postag="Sp*", because we search for another subordinate word, which is the head of a subtree.

A disadvantage of the UAIC annotation convention is that there is no information about the syntactic relationship of the relative pronouns and adverbs in the subordinated sentence, but only about their role in the complex sentence. Therefore, they can only get the tag mark, obtained by automatically changing the subord. relationship.

A human annotator must specify their function in the subordinate sentence, because the relative pronoun, adjective and adverb are not marks. Sometimes the relative pronoun is a nominal modifier in the subordinate sentence, and then its automatic subordination to the head of the subordinate sentence by the transformation 2 is erroneous and will be manually corrected.

5.2 Subordination of the Copulative Verb and of the Subject to the Predicative Noun

This transformation has the aim of swapping the places of the copulative verb and the predicative noun, in the cases where the copulative verb is the verb “to be”. In the UAIC conventions of annotation, there are other 9 verbs annotated as copulative. Their predicative nouns becomes xcomp in the UD convention.

This transformation has the advantage that it establishes an equivalence between the structure of dependencies of the nominal predicate and the passive voice, which is also built with the verb “to be”, subordinated to the verbal participle, which formally resembles an adjective predicative noun (and it has the same number and gender with the subject).

This transformation is as follows: if there is a word with id="x", and lemma="fi", and postag="Vm*" and head="q" (and if there is a word with id="y", head="x", and deprel="sbj."), then the word with deprel="n.pred." changes head="q" and the word with id="y" changes head="z". The deprel="n.pred." is changed with the deprel of the word with id="z", and this one take the deprel CNCOP.

A general condition must be observed for all the rules: If the id of the head changes, then all the other words having the changed head must also change their head. Therefore, the dependencies of the copulative verb “to be” become dependencies to the predicative noun, i.e. they change the head="x" in head="z". For this rule, the exception are the words with morphological analysis (POSTag) Qn, aux., Qs.(see Figure 2).

5.3 Subordination of all Coordinated Elements to the First One

In the UAIC convention, the coordination is rendered by an asymmetrical tree, also having as head the first element, but each coordinating element (word with full meaning, functional word or punctuation) being subordinated to the element above and acting as the head of the element below. Similarly with the subordinated elements of relation considered as heads, the coordination elements also are heads and are positioned between the related elements.

Consequently, in the new conception, the relational words for the coordination must be subordinated to the meaningful words which they introduce, by a rule similarly formulated as the first rule in the 5.1 chapter.

Then, all the coordinated meaningful words must be subordinated to the first one. A rule for retaining the previous head is added, because in a sentence with more coordination relations they must be subordinate to the first element of their chain, and not to another one.

The elements with the deprel="coord.", and the
postag="COMMA", are also subordinated to the meaningful word which they coordinate. Simultaneously, the deprels of all these words change, differently in UD and in the Semantic annotation.

The result of this transformation is shown in Figure 3

As can be seen, in Figure 3 and 4 the transformation of relational words and of the copulative verb has already been made.

Coordination in sentences takes place between long-distance elements. Because it acts in the top of the tree, this transformation must be applied after all the others.

In Figure 4, the following text is annotated in two conventions:

“I think that there will come some days of effort that will pass, that we will escape and that we will be happy.”

The results of applying the other transformations can be seen in the figure 4 a: First, the preposition de “of” has been subordinated to the noun “effort”. Secondly, the conjunction “că” (“that”), repeated three times, and the relative pronoun “care” (“which”) have been subordinated to the head of the subordinate sentences. Then, the predicative noun “fericiţi” (“happy”) is the head for the copulative verb.

Finally, the “ccomp” sentences 2 and 3 have been subordinated to the first one, and the comma and the conjunction “şi” (“and”), the two connectors of the three coordinate sentences, are subordinated of the closest coordinated element on the left. (see Figure 4).

The rules for the transformations have been all formulated in the same way of the listening in the 3.1 chapter.

6 Conclusions and Future Works

Transforming UAIC-RoDepTb so that it can be used in multiple future applications and can be compared to similar corpora is one of our priorities.

TREEOPS, the program described in this paper (section 4) is a very important tool. It is language independent. For any resource in XML format to be transformed, another set of rules can be written, depending on the original format and on the one in which it is intended to be transformed, regardless of the language. As we have shown above, format flexibility is very important for all resources, so they can always be compatible with newer resources created in more modern formats. The TREEOPS program does not have a variant for CONLLU, but we have transformation programs from XML to CONLLU and vice versa. TREEOPS has been successfully tested to transform the classic format of our UAIC treebank into the semantic format, we will begin testing TREEOPS for the conversion from classical UAIC to the UD format; after completion, the program will be available on the Sourceforge page. 5 The evaluation of the accuracy is in progress. It is difficult to evaluate it, because the program does not transform all the semantic structure, but only the non ambiguous relation; the manual transformation of the ambiguous one should not be evaluated as a decrease in the accuracy of the conversion program.

Another important task is to ensure optimal digitization of the old Romanian language information, starting with its first attestations. Digitization does not only mean scanning old manuscripts and prints to avoid their disappearance with paper damage, but also reading the data contained in them.

For this, a very useful tool has been created and, for the first time ever, seventeenth-century

5https://sourceforge.net/
Cyrillic Romanian letters have been made editable by an optical character recognizer (OCR) built in Chisinau (Colesnicov et al., 2016), (Cojocaru et al., 2017).

Another important tool is the OldRoPOSTagger (Măranduc et al., 2017b), which provides the first level of annotation of the texts obtained after their transliteration into Latin characters. In order to syntactically parse these texts with many particularities, especially related to the very free order of the words and subordinate sentences, a continuous training of the syntactic parser is needed.

By creating the converter described in this article, we will transfer all these sentences in the UD and Semantic format. The operation is under way. Transformations will be supervised and the rules for the conversions improved.

We also aim to include the oldest New Testament (Alba Iulia 1648) in the project Pragmatic Resources in Old Indo-European Languages (PROIEL), that is part of the Universal Dependencies (UD) group, and align the oldest New Testaments in Latin, Greek, Armenian, Slavonic. The philological studies of old translations and the etymological studies can benefit from the alignment of the first printed Romanian New Testament in this project. The first part of the New Testament (1648), the Gospel, was introduced in the classic syntactic format and in the semantic one. We also have to introduce the second part, the Acts of the Apostles.

Since the UAIC-UD transformation does not require the introduction of new syntactic relationships, we do not believe that we need to build a syntactic parser on the UD format. A parser for Contemporary Romanian in UD format has been created (Mititelu and Irimia, 2015), and we can train it on Old Romanian, too.

Once a large training corpus has been built in the semantic format, we will create a semantic parser. The semantic parser will complete the set of tools for processing old Romanian.

**References**

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. *The berkeley framenet project*. In *Proceedings of the COLING-ACL*. Montreal Canada.

Eduard Bejček, Eva Hajičová, Jan Hajič, P. Jínová, Vaclava Kettnerová, Veronika Kolářová, Marie Mikulová, Jíří Mírovský, Anna Nedolužko, Jamila Panevová, Lucie Poláková, Magda Ševčíková, J. Štěpánek, and Sarka Zikánová. 2013. *Prague Dependency Treebank 3.0. Data/software*. Univerzita Karlova v Praze, Prague.

Alena Bohmová, Jan Hajič, Eva Hajičová, and Barbora Hladka. 2003. *The Prague Dependency Treebank: A Three-Level Annotation Scenario. Text, Speech and Language Technology*. Springer Publisher, Prague.

Laura Bănărescu, Claire Bonial, Shu Cai, Mădălina Georgescu, Kira Griffith, Ulf Hermjakob, Kevin Knight, Philip Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation
for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. pages 178–186.

Svetlana Cojocaru, Alexander Colesnicov, and Ludmila Malahova. 2017. Digitization of old romanian texts printed in the cyrillic script. In Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language. Alexandru Ioan Cuza University Press, pages 1–11.

Ioachim Druguș. 2015. Metalingua: a metalanguage for the semantic annotation of natural languages. In Proceedings of the 11th International Conference Linguistic Resources and Tools for Processing the Romanian Language. Alexandru Ioan Cuza University Press, pages 79–94.

Florinel Hociung. 2016. Treebank Annotator - dissertation. Faculty of Computer Science, Alexandru Ioan Cuza University, Iași.

Franz Kern. 1883. Zur Methodik des deutschen Unterrichts. Nicolaische Verlags-Buch-handlung, Berlin.

Cătălina Mărănduc and Cenel-Augusto Perez. 2015. A romanian dependency treebank. International Journal of Computational Linguistics and Applications 6(2):25–40.

Cătălina Mărănduc, Monica Mihaela Rizea, and Dan Cristea. 2017a. Mapping dependency relations onto semantic categories. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).

Cătălina Mărănduc, Radu Simionescu, and Dan Cristea. 2017b. Hybrid pos-tagger for old romanian. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Ldc.upenn treebank 3. In Technical Report. University of Pennsylvania, pages 1–230.

Igor Mel’čuk. 1988. Dependency Syntax: Theory and Practice. The SUNY Press, Albany, NY.

Verginica Mititelu, Cătălina Mărănduc, and Elena Irimia. 2015. Universal and language-specific dependency relations for analyzing romanian. In Proceedings of the Third International Conference on Dependency Linguistics, Depling. Uppsala University, pages 28–37.

Verginica Barbu Mititelu and Elena Irimia. 2015. Types of errors in the automatic syntactic parsing of romanian. In Errors by Humans and Machines in multimedia, multimodal and multilingual data processing. Proceedings of ERRARE 2015. Sinaia, Romania, pages 195–204.