RoDia Corpus: Fostering Language Diversity in One Corpus

Victoria Bobicev
Technical University of Moldova
St cel Mare bvd. 168
Chişinău, Moldova
victoria.bobicev@ia.utm.md

Cătălina Mărănduc
''Iorgu Iordan – Al. Rosetti'' Linguistics Institute
Calea 13 Septembrie nr. 13,
Bucureşti, România
catalinamaranduc@gmail.com

Relevant UniDive working groups: WG1

1 Introduction

We present a corpus with rich morphological, syntactic and partially semantic annotation. Its main characteristics are the large variety of non-standard texts and several types of annotation.

The creation of this corpus pursues several objectives: (1) a better coverage of linguistic diversity of Romanian language; (2) diachronic analysis of Romanian; (3) creation of a gold standard annotation for various types of Romanian texts which permits (4) creation of robust machine learning models for various types of annotation.

2 RoDia Corpus

The corpus we are working with currently is the version of Alexandru Ioan Cuza University (UAIC) Romanian Dependency Treebank (UAIC-RoDia-DepTb) (Mărănduc et al., 2017a). It has been started by Catalina Mărănduc as non-standard corpus and it is presented in several formats: (1) the syntactic classic, following the rules of syntactic annotation developed at UAIC; (2) the UD syntactic, and (3) a new syntactic-semantic one. At the current moment, it is the biggest syntactically annotated Romanian corpus[1]. The UD Romanian-nonstandard treebank is based on UAIC-RoDia Treebank (UAIC-RoDia-DepTb) with its rich morphological and syntactic annotations. The annotation of UAIC-RoDia Treebank has been transformed in UD conventions and uploaded on the UD page.

2.1 Corpus Annotation

RoDia (Romanian Diachronic) corpus contains non-standard types of texts (Malahov et al., 2017). The standard language is rarely used in human communication: official relationships, scientific reports, books for publication, exams. Simplified standard examples give little information about linguistic creativity. This is the reason we concentrate on the annotation of non-standard text types such as oral regional fiction, social media communication, poetry, historical Romanian texts and others. We would like to cover all types of texts' variations: diatopic, diastratic, diamesic, diaphasic and we have made some steps in this direction adding to our corpus modern chat texts and Moldova's folklore.

The whole corpus has morphological and syntactic annotation using the dependency grammar conventions. The annotated files are coded in XML format.

POS-tags used in UAIC morphological annotation have been developed by MULTEXT project (Erjavec, 2012) for several languages in order to create universal type of morphological codes that could be used for as many languages as possible especially morphologically rich ones. 614 morpho-syntactic tags have been created for Romanian due to its rich morphology. In practice, the number of tags used for annotation has been slightly reduced; however it is still more than 500.

44 syntactic labels are used for the dependency relations. Examples of syntactic labels are: sbj. (subject), c.d. (direct object) c.c.t. (circumstantial temporal modifier of verb). The main predicate is coded as root and the other words are dependent on it. Semantic annotation is also coded in xml and has similar format, except the dependency relations (deprel), which present 96 types of semantic relations (Mărănduc et al., 2017b). 20 of the syntactic tags have a single semantic correspondent; the other 24 have several semantic interpretations and correspond to 74 semantic tags. Semantic interpretations have been added mostly to the modifiers . 13 of the 14 circumstantial modifiers are equivalent syntactic and semantic tags, except the modal circumstantial, which is ambiguous and has several possible semantic interpretations such as Comparative, Intensifier, Restrictive and several other labels. Various semantic interpretations have been added to noun's modifiers, for example: "Bani pentru excursie" (Money - for the trip) "Pentru excursie" is a nominal modifier with a purpose meaning - PURP.

[1] https://universaldependencies.org/

Nr.	Format	Sentences	Tokens
1	UAIC syntactic XML	32,753	671,235
2	UD syntactic CoNLLU	26,225	572,436
3	UAIC semantic XML	5,566	99,341

Table 1: Volume of the corpus annotated in each of the three formats: UAIC syntactic, UD syntactic and UAIC semantic.

One of the similar projects is Prague Dependency treebank (Bejček, 2012). The Czech researchers refer to this treebank as a three-level annotated corpus of 1.8 mil. tokens. The first level is the morphological annotation; the second is the superficial syntactic annotation, in the UD annotation conventions, and the third one is called the tectogrammatical level, or the level of linguistic meaning. The semantic and syntactic annotation of our corpus have closer relations. Semantic one is more detalied description of the syntactic relations.

The process of the annotation is a classical one. Firstly, the texts are pre-processed. In some cases the pre-processing is a complex task as, for example, in the case of old Romanian texts printed in Cyrillic described in (Cojocaru et al., 2017). The OldRo-POS-tagger (Mărănduc et al., 2017c) is used for PoS tagging. It was obtained by introducing numerous old and regional variants of specialized dictionaries into the POS-tagger lexicon.

For syntactic annotation, we tested MaltParser which was considered the basic one for UD corpora on our corpus. We evaluated nine parsing algorithms of the parser. Training on train part of our UD corpus and testing on the testing part we reached the acuracy of 75% for LAS (Labelled Attachmnet Score) and 86% for UAS (Unlabelled Attachment Score) (Mărănduc et al., 2020). We understand that modern parsers based on deep learning models can achieve much higher accuracy and our future plans are to experiment in this direction. However deep learning models need large volumes of training data and it is yet to be found whether our corpus is large enough to train such kind of parser.

Both morphologic and syntactic annotations are checked manually by the experienced linguist. This allow us to say it is a gold standard annotation although only one person worked on the annotation verification and correction.

2.2 Comparison of UAIC and UD Formats

In order to make our corpus widely accessible for the reserch community we transformed it to Universal Dependencies (UD) format which is required to be done strictly following the UD instructions. The common format makes possible alignment of different languages and automate translation or comparative language studies.

UD annotation is coded in CoNLLU format[2] and UAIC in xml but this was not the main problem for the transformation. UAIC and UD dependency structures have different structural principles. The UD annotation convention highlights words with full meaning and the connecting words (prepositions, conjunctions etc.) are subordinated to them. In the UAIC convention, the connecting words are the heads. In the UD system, it is easier to compare texts in very different languages and to emphasize the semantic structure. In the UAIC system, the logical structure consisting of semantic units and the function words that play role of connectors are the nodes to which the words with the main meaning are connected. UAIC format have 14 kinds of circumstantial modifiers, which is a rich source of semantic information that we could not loose. A specific type of annotation called semantic has been created in order to preserve the semantic information already annotated in UAIC format (Mărănduc et al., 2018).

Although there were graph transformation tools, for example, (Guillaume Bonfante, 2018), we did not find any at a time and a rule-based transformation tool called TREEOPS has been created in order to transform UIAC annotation format in UD (Bobicev et al., 2017). The tool uses a customized set of rules of several types: the simplest ones which change only one tag on another tag; more complex ones which have several conditions; and even more complex ones which change the tree structure. Two separated sets of rules were created for UAIC: (1) UAIC -> UD and (2) UAIC -> semantic format. It should be pointed out that in

[2]https://universaldependencies.org/format.html

the semantic format, some transformations are ambiguous. This is why the automate transformations need to be completed by human annotators.

2.3 Current Statistics for our Corpora

The current volumes of the corpora in three formats described above are shown in the Table 1. In spite of the impressive number of manually verified sentences with syntactic annotation we still lack of good model for the reliable automate syntactic annotation. This is one of our main concern and our aim is to find a robust tools which could produce a qualitative annotation that would require minimal human supervision.

3 Conclusion

The main aim of our work is the creation of the gold standard corpus to be used for future training of part of speech taggers and syntactic parsers; its volume should be enough for reliable parsing with minimum errors.

On the other hand, we need good annotation tools for faster corpus creation. Thus, our goals are interdependent: the corpus creation is dependent on the tools and the tools need a corpus for their training.

Until now, we worked iterative, adding small manually corrected parts to the main training corpus and re-training our parser each time.

Given the rapid progress in language technology we believe that we can find and adapt a pipeline of tools that could help us expand our corpus faster and include a wider variety of documents in the corpus.

References

Panevová J. Popelka J. Straňák P. Ševčíková M. Štěpánek J. Žabokrtský Z. Bejček, E. 2012. Prague dependency treebank 2.5 – a revisited version of pdt 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246.

Victoria Bobicev, Cătălina Mărănduc, and Cenel Augusto Perez. 2017. Tools for building a corpus to study the historical and geographical variation of the Romanian language. In *Proceedings of the First Workshop on Language technology for Digital Humanities in Central and (South-)Eastern Europe*, pages 10–19, Varna. INCOMA Inc.

S. Cojocaru, A. Colesnicov, and L. Malahov. 2017. Digitization of old romanian texts printed in the cyrillic script. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, page 143–148, New York, NY, USA. Association for Computing Machinery.

Tomaž Erjavec. 2012. Multext-east: Morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, 46.

Guy Perrier Guillaume Bonfante, Bruno Guillaume. 2018. *Application of Graph Rewriting to Natural Language Processing*, volume 1. ISTE Wiley, London.

Ludmila Malahov, Cătălina Mărănduc, and Alexandru Colesnicov. 2017. A diachronic corpus for Romanian (RoDia). In *Proceedings of the First Workshop on Language technology for Digital Humanities in Central and (South-)Eastern Europe*, pages 1–9, Varna. INCOMA Inc.

Cătălina Mărănduc, Victoria Bobicev, and Roman Untilov. 2020. Parsing romanian texts. In *2020 13th International Conference on Communications (COMM)*, pages 331–334.

Cătălina Mărănduc, Hociung Florinel, and Bobicev Victoria. 2017a. Treebank annotator for multiple formats and conventions. In *Proceedings of the Conference of Mathematical Society of the Republic of Moldova*, pages 529–534, Chisinau. Instrumentul Bibliometric National.

Cătălina Mărănduc, Cătălin Mititelu, and Victoria Bobicev. 2018. Syntactic semantic correspondence in dependency grammar. In *Proceedings of the 16th International Workshop on. Treebanks and Linguistic Theories*, pages 167–180, Prague.

Cătălina Mărănduc, Monica-Mihaela Rizea, and Dan Cristea. 2017b. Mapping dependency relations onto semantic categories. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, Budapest.

Cătălina Mărănduc, Radu Simionescu, and Dan Cristea. 2017c. Hybrid pos-tagger for old romanian. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, Budapest.