# Non Standard Treebank Romania – Republic of Moldova in the Universal Dependencies

Cătălina Mărănduc, Victoria Bobicev

**Abstract**

Morphological and syntactic annotated corpora are very important for language computerization. A Dependency Treebank was created in 2007 at the Al. I. Cuza University. 4,500 sentences of it were introduced into the Universal Dependencies (UD), consisting of over 80 treebanks for 50 languages annotated in unique conventions. Added to other 5,000 sentences annotated at the Artificial Intelligence Institute in Bucharest, they form a Romanian Contemporary Standard Treebank, affiliated with UD. But our Treebank has other non-standard language sentences. We intend to affiliate with UD a Non-Standard Romanian Dependency Treebank, having 15,000 sentences, part of them collected and annotated in the Republic of Moldova.

**Keywords:** treebank corpus, dependency grammar, syntactic parser, non-standard language, language specific peculiarities.

## 1 Introduction

The Al. I. Cuza University Romanian Diachronic Treebank, UAIC-RoDia, (ISLRN 156-635-615-024-0) has now 16,187 sentences, with 322,404 words, punctuation included. The 4,500 sentences in Contemporary Standard Romanian, which were affiliated with UD, come from the research conducted by Augusto Perez within the doctoral thesis and are mostly translations from English [12].

Initially, the UD project aimed to create a universal syntactic parser, and for this purpose, not very complex sentences in Contemporary

Standard languages are requested. The general features were highlighted and the language specific ones were admitted only as sub-classifications.

However, as the group grew up, other uses of affiliated treebanks emerged, comparative language study, old language study, or automated translations. All types of treebanks were allowed, both consisting of a single type of text and balanced ones, like ours, with texts from social media communication to old texts.

In fact, there are few occasions when standard language is used in communication, official relationships, scientific communications, books for publication, exams. We cannot study and process the natural language only on the basis of the simplified standard examples; especially they do not give information about linguistic creativity. That's why we've annotated more and more non-standard text types: oral regional fiction, social media communication, poetry.

## 2 Related Work

At the moment, the international community is interested in preserving and digitizing the cultural heritage, i.e. the processing of old texts. One of the projects affiliated with UD is PROIEL (Pragmatic Resources in Old Indo-European Languages) It contains the New Testament in Latin, Ancient Greek, Ancient Slavonic, Aramaic languages. We choose to introduce in the UAIC Treebank the Alba Iulia New Testament (1648), the first printed in Romanian, with the intention to compare it with the other Old New Testaments in this project [8]. We have already annotated the four Gospels and we are going to annotate the Acts of the Apostles.

The text with Cyrillic letters was obtained using an Optical Character Recognizer (OCR) made at the Institute of Mathematics and Computer Science in Chisinau [4]. We provided data from our corpus for the old language lexicon required for this OCR program.

Participating at the DATeCH (Digital Access of Textual Cultural Heritage) conference in Gottingen on June 1-2, 2017 (https://www.digitisation.eu/datech-international-conference), we noticed that our OCR program for old Romanian Cyrillic letters and Part of Speech (POS) tagger for Old Romanian are compatible with similar programs presented, for example, OCR for the Old Gothic letters [6].

More UD-affiliated treebanks have another format outside the UD one, for research on the language specific peculiarities, semantics, pragmatics, text annotation, which the international format does not favors. For example, the Tectogrammatic layer of Prague Treebank [2], or the Head-Driven Phrase Structure Grammar (HPSG) format of the Bulgarian Treebank [11].

## 3 The Regional Folk Poetry

Oral folk creation is also part of the cultural heritage and must be preserved and protected, especially as a phenomenon of extinction as the written culture spreads in the villages.

Computer scientists who prefer simple texts to get a good accuracy of processing tools have always avoided the annotation of the lyrics. But these have also to be taken into account as a creative phenomenon of natural language. Recently, a project lead by Cristina Vertan has begun in Hamburg with the purpose to annotate poetry.

Another topic of our research is to compare the language spoken in Romania to the one spoken in the Republic of Moldova. Lexical studies were recently conducted on journals published in the two countries in the nineteenth century and at the beginning of the twentieth century. Their results show that the differences between them are minimal [7].

However, the differences appear at the topic and syntactic level, as well as in the non-standard language used by villagers who do not have access to normative Romanian texts.

A recent study on social media communication in the Republic of Moldova shows that many Russian words are used in non-standard language. This is a peculiarity of the bilingualism. The same study shows that speakers control these linguistic interferences and exclude them when a person who does not speak Russian participates in the conversation [5].

Both folklore and lyrics with rhymes are poorly processed and annotated by Natural Language Processing (NLP) specialists around the world. For all these reasons, we decided to make a comparative study of popular texts with rhymes in Romania and the Republic of Moldova. We intend to organize this annotation on the UD platform for the international visibility of this study. In Romania, we started to annotate local texts from

Muntenia and Oltenia. At the same time, an annotation of a collection of Moldovan Ballads began in the Republic of Moldova [1].

By now, we have been annotating in XML format and in UAIC treebank conventions, because we do not have any processing tools for another format. A large gold corpus in the new format is necessary in order to build the necessary tools. A tool called Treeops was created to transfer XML from one format to another desired format [3]. Using this tool, we plan to transform the texts in the UAIC format into the Universal Dependencies one.

It should be pointed out that these transformations need to be corrected by human annotators. We have two work interfaces to do it [9]. We also have converters from the XML format to the CONLLU format used by UD. This transformation does not require supervision and is performed in automate mode.

## 4 Brief Introduction to Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 50 languages [11].

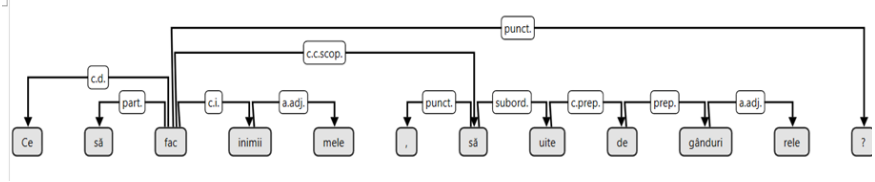

Figure 1. An example of an annotated sentence in UAIC format.

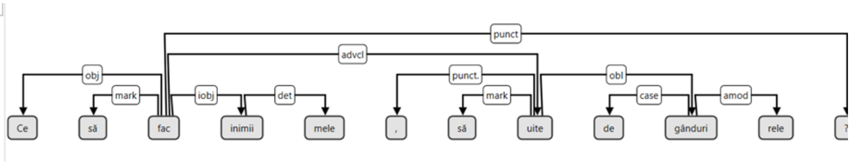

Figure 2. An example of an annotated sentence in UD format.

These are treebanks built into the dependency grammar system: the words and the punctuation are nodes, and the arcs of the graph are inscribed with the relationships between them; no equality is allowed, but only subordination; the subordination from more than one head is no allowed [13]. This is an economical and flexible system similar to finite state automata.

The UD annotation convention highlights words with full meaning and the relational words are subordinated to them (see Figure 2). In the UAIC convention, the related words are heads (see Figure 1). 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 connectors are more visible.

## 5 Conclusion

The paper presents an ongoing work of syntactically annotated corpora creation. Several efforts were made to enrich the standard corpora with non-standard and more difficult annotated examples such as folklore and lyrics. Sub-corpora from different regions where Romanian language is spoken are in the process of creation and annotation. The necessary volume of annotated and manually corrected texts would serve as a training corpus for the statistical parser.

A resource is the more useful as while it is enriched and has a flexible form, easy to adapt to international formats. Our affiliation to UD will create a great visibility of our common efforts and perspectives to participate in international projects.

## References

[1] V. Bobicev, T. Bumbu, V. Lazu, V. Maxim, D. Istrati *Folk poetry for computers: Moldovan Codri's ballads parsing.* Proceedings of the 12th International Conference "Linguistic Resources and Tools for Processing the Romanian Language (2016), pp. 39-50.

[2] A. Bohmova, J. Hajic, E. Hajicova, B. Hladka. *The Prague Dependency Treebank: A Three-Level Annotation Scenario*. Text, Speech and Language Technology. Springer Publisher, Prague. (2003).

[3] M. Colhon, C. Mărănduc, C. Mititelu, *A Multiform Balanced Dependency Treebank for Romanian.* Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities, (KnowRSH), pp. 9-18,

[4] S. Cojocaru, A. Colesnicov, L. Malahov. *Digitization of Old Romanian Texts Printed in the Cyrillic Script.* Proceedings of DATeCH (2017), pp. 143-148. https://www.digitisation.eu/datech-international-conference/

[5] V. Cojocaru. *Discourse markers in Romanian spoken in the Republic of Moldova: pragmatic and sociolinguistic aspects*. PhD Thesis, Faculty of Letters, University of Bucharest. (2016).

[6] F. Fink, K. U. Schulz, U. Springmann *Profiling of OCR'ed Historical Texts Revisited*. Proceedings of DATeCH (2017), pp. 61-66.

[7] D. Gîfu. *The Analysis of Diachronic Variation in Romanian Print Press*. In: Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems, NESSUS PhD (2016), pp. 49-53.

[8] D. T. T. Haug. *The PROIEL corpus: annotation of morphology, syntax and information structure*, Perspective Project kick-off meeting, University of Nijmegen, (2014).

[9] C. Mărănduc, F. Hociung, V. Bobicev, *Treebank Annotator for multiple formats and conventions*. Proceedings of The 4th Conference of Mathematical and Computer Science Society of the Republic of Moldova, (2017), pp. 529-534.

[10] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman. *Universal Dependencies v1: A Multilingual Treebank Collection*. Proceedings of LREC (2016). http://universaldependencies.org/

[11] P. Osenova, K. Simov. *Syntactic-Semantic Treebank for Domain Ontology Creation*. Cognitive Studies. SOW Publishing House, Warsaw, Poland, (2011), pp 213-225.

[12] C. A. Perez. *Linguistic Resources for Natural Language Processing*. PhD thesis. Faculty of Computer Science, Al. I. Cuza University, Iasi, (2014).

[13] P. Tapanainen, T. Jarvinen. *Towards an implementable dependency grammar*. CoLing-ACL98 workshop Processing of Dependency-based Grammars. (1998).

Cătălina Mărănduc[1,2], Victoria Bobicev[3]

[1]Faculty of Computer Science, Al. I. Cuza University, Iași
catalinamaranduc@gmail.com:

[2]Iorgu Iordan – Al. Rosetti Academic Institute of Linguistics, Bucharest

[3]Tehnical University of Moldova
victoria.bobicev@ia.utm.md: