“Alexandru Ioan Cuza” University of Iași, Romania
Faculty of Computer Science

INSTRUMENTS FOR PROCESSING THE OLD ROMANIAN LANGUAGE (Training Corpora, Models, Statistics)

Summary

PhD candidate
CĂTĂLINA MĂRĂNDUC

Scientific advisor
DAN CRISTEA

Contents

For a long time, modern linguistics had a rather limited point of view over the object of study: we are interested on the contemporary language, mainly because it is directly accessible to us, we are in direct contact only with it, and anything else is of very little interest. Therefore, we cannot be surprised that this was also the point of view of computer scientists who dealt with natural language processing, at least until the end of the last century. Moreover, as the domain of computational linguistics matures and it started to have a history, problems appeared. Some are related to the storage of information from the proceeding research. As programs that process linguistic data evolved so fast, old data could no longer be read by the new programs. Even more, storing linguistic data from past centuries revealed to be a difficult task, since computers must be able to access this information and allow advanced searches onto these data.

In the early phases of building automatic morphological and syntactic analysers, the purpose was to be able to process only simple phrases, so difficult constructions were deliberately eliminated from training corpora. The very last aim of our enterprise is, nevertheless, to have a technology able to analyse any phrase of the Romanian language, not only simpler ones.

In this thesis, we are not interested in building a big corpus of contemporary standard language. A large corpus for the standard Romanian language has already been built. Its name is CoRoLa[1] (Barbu-Mititelu *et al.*, 2017), and it was launched in November 2017, having in its first delivery almost 400,000 files, around 1.26 billion

[1] http://corola.racai.ro/

tokens (words + punctuations), and approx. 900,000,000 word occurrences. This corpus is entirely morphologically annotated and is undergoing automatic syntactic annotation.

Consequently, our purpose in this thesis is to build corpora for the nonstandard Romanian: a chat corpus, a corpus for the old Romanian language, and a corpus for the regional Romanian language, spoken in Moldova on both sides of the river Prut. Then we will train various processing tools on the created corpora and adapt them so that we get optimal results in automatic morpho-syntactic analysis. A corpus of Romanian Contemporary Standard is not representative of how the natural language really looks like. The regional texts, poetic texts, the spoken language, familiar language, the language of social media, journalistic texts which can make use of poetic images or specific means of oral-familiar language, all these types of language frequently contain non-standard phenomena. They must also be automatically processed.

The aim of our research is to create a wide and balanced corpus of nonstandard variants of the Romanian language, with special attention on the Old Romanian language, and to study how various natural language tools can be trained on each of the language variants for which we have succeeded to have a sufficient training corpus. We will see what the particularities of each such variant of the language are, what errors appear in the processing, what their causes are and how we can optimize the processing tools and increase their performance.

The structure of this thesis is as follows. We present in chapter II the current content and extent of training corpora UAIC syntactic, UAIC semantic, UD syntactic. The transformation programs and their accuracy will also be described: the two variants of Treeops and the program for the transposition of the XML into CONLLU format of the UD treebanks.

Then, in Chapter III, we explain the sets of morphological tags for each convention, and the morphological layer of annotation. We also describe a tentative to create a POS-tagger for the Old Romanian (Mărănduc *et al.* 2017), based on the UAIC hybrid POS-tagger (Simionescu, 2011). This tool is still in work, because a training corpus of at least 60,000 sentences is needed.

In Chapter IV we present the Syntactic conventions of annotation of two corpora, the differences between them, the tools for the transposition to the basic format into US one, and syntactic parsers applied on both conventions, and statistics.

In Chapter V, the semantic convention is presented, and the rules for the transposition of the basic format into the semantic one.

In Chapter VI, we describe the operation of extracting structures for the elaboration of the Pattern Dictionary of Romanian Verbs, this being only at the beginning of the elaboration. We show the types of tree accesses, in search, into the developed corpora.

In Chapter VII an application is presented, aligning the New Testament from Alba Iulia 1648 with the Greek, Latin and Old Slavonic versions of the holy book. These versions represent for potential specialists sources of the Romanian translation and their comparative study is very important to clarify the origin in the old Romanian language of words and structures.

The concluding chapter VIII makes a brief summary of the original contributions of the thesis and indicates the directions in which the created resources must be further developed and ways in which they can be used in the research of linguists and computational linguists.

In the following we will summarise each chapter in turn.

The **first chapter** gives an introduction to the work presented in the thesis and gives a detailed description of the steps taken in achieving our corpus.

Chapter II presents the training corpora and our treebanks. The UAIC training corpus includes now **32,753** sentences, and **671,235** tokens. But a syntactic parser for the Romanian language should be trained separately for each category of texts. The chat and the old language are very different from a syntactical point of view from the current language. As such, we separately trained a Malt parser for social media texts. The parser, trained on a sub-corpus of only 2,579 sentences, achieved better results after adding all the contemporary texts (Perez *et al.*, 2016 a, b). Therefore, both the contemporary and the specialised (old, social, etc.) sections should grow to improve the accuracy of a fully automatic process of syntactic annotation. With the current technology and respecting the same rules, the minimal limit for the Contemporary Standard language is 10,000 sentences. For a category of texts in which a variation of grammatical rules applies over time, because of the evolution of the language, the number of sentences should be bigger. Social media texts and the oral folklore, being creative, assume very much freedom regarding grammar rules. Therefore, we need also more than 10,000 sentences for each of these registers. For instance, for an optimal training of the POS-tagger, we estimate the minimal size of the training corpus to be 70,000 sentences.

What we have accomplished so far is a considerable corpus in three formats. In Table 8, the whole content of our database can be seen.

Table 8. The content of all the formats of our Treebank

Nr. rt.	Format	Sentences	Tokens	Average tokens/sentence
1	**UAIC syntactic XML**	**32,753**	**671,235**	**20.49**
	from which, Old-Ro	19,254	126,564	21.47
1a	**word-form Cyrillic**	2,794	46,708	16.71
2	**UD syntactic**	**16,936**	**348,562**	**20.58**

	CoNLLU			
	from which, Old-Ro	14,437	297.109	20.57
	from which, Folklore	2499	50,077	20.03
3	**UAIC semantic XML**	**5,566**	**99,341**	**17.84**
	from which, Old-Ro	5,032	88,350	17.55
	from which, folklore	230	4,157	18.07
Tot	**corrected**	**55,255**	**1,119,138**	**20.25**

In **chapter III** we show that our purpose is not as much to grow corpora in the tree convention of annotation, as is to train on these types of corpora more tools for processing Old Romanian. Any future application based on these corpora would be imagined, and in any kind of convention, there is no one that does not have to base the segmentation of texts into words and their morphological annotation. Thus, the automatic morphological annotation is the basis of more advanced natural language processing and we must begin with it. To build a POS-tagger for Old Romanian, having a similar tool trained on Contemporary language, it is necessary to, first, establish the list of tags, then to add to the POS-tagger lexicon word forms of the ancient language, and, finally, to have a training corpus with a big number of sentences consistently annotated with the new tag set.

Syntactic information (**chapter IV**) is essential for natural language processing. The input string is first decomposed by programs such as the splitter and the tokeniser, then segments are analysed by the POS-tagger and, finally, syntactic information is recomposed into a structure. Each token has a head, except the root of the sentence. The relationship by which it binds to its head can be optional or mandatory and is of several types. For each of these places determined in the structure, elements with certain morphological characters are chosen. Syntax is related to both morphology and semantics. In this chapter we present different constraints met in our process when annotating coordination, predicative nouns or double roles.

Our purpose is to introduce in the UD-Romanian-nonstandard other Old Romanian language texts. UAIC treebank has 6,590 sentences, with 162,231 words and punctuation marks in work. Besides Old Romanian and Folklore, the Chat corpus will be inserted in the UD-Romanian-Nonstandard treebank. We will continue to edit the Pattern Dictionary of Romanian Verbs; it will have a site, where the patterns of the Old and Regional Romanian will be marked. These patterns can be used to create a new syntactic parser or a mixed semantic-syntactic model based on constraints, permissions and bans.

In **chapter V** we present details about the semantic annotation. We have made the transposition table for automatically transposing our conventions into UD ones and part of the UAIC-RoDiaTb was transposed in the UDV in 2016, by the RACAI group (Research Institute of Artificial Intelligence). There are many theoretical problems that differentiate the conventions of UAIC from the UD ones; for example, the treatment of relational words. The syntactic categories are classified according to the UD conventions in what concerns the morphological classes (i.e., adjectival, adverbial, nominal modifier); additionally, we consider that the syntactic information should be correlated with the semantic one.

This chapter proposes a type of semantic annotation with more categories, since we aim to keep all the information that has been annotated in the UAIC syntactic layer. This information is important since it can be exploited by other applications. Another purpose was to find an international standard of annotation with similar categories, in view of a future affiliation. The similarities with the tectogrammatic layer of PDT and with the AMR logical categories are obvious. However, there are also differences since the resultant graph of the AMR semantic annotation is not a dependency tree, and the nodes are not words, but concepts. In order to show the isomorphism between the syntactic and the semantic structures, we chose to build a corpus of semantic dependency trees, with similar tags to the tectogrammatic layer of the PDT.

We also discussed the transformation process of UAIC RoDia Dependency Treebank syntactic annotation into the logical-semantic annotation. This transformation is done automatically for non-ambiguous syntactic relations, and manually for ambiguous relations. In the future, we aim to transform the syntactic and morphological annotation of the second part of the New Testament of 1648 (*Acts and Letters of the Apostles*) into a semantic annotation. We will train a statistical parser on this corpus, in order for the parser to learn to transform ambiguous syntactic relations.

Chapter VI presents the availability of our resources. The UAIC RoDia Dependency syntactic corpus resides in the XML format and includes facilities for advanced tree search. This treebank is open source. The site will include links to the Romanian language UD corpora, in CONLLU format, which are also open source. To allow the user to download data in the preferred format, XML-CONLLU converters and vice versa will be included.

A web site is the definitory location of a resource. It cannot be found and used by the interested persons if it has not a web address. Each language in the big UD treebanks family, contributed by more than 150 research teams spread all over the world, uses the same basic representation format, the specific treebanks being each located somewhere in the virtual space. Out of the specific annotation conventions of the contributing languages, the UD form is abstracted as a common international convention and, for each contributing treebank, this form is obtained either through a semi-automatic or a completely automatic transformation process.

All resources are linked, aligned, compared, and integrated in the large UD family. Authors, contents and original web sites are disseminated through descriptive articles, which present also results of the activity of the creators.

In **chapter VII** we present the alignment of *The New Testament* and its conventions of annotation. The alignment is useful to translations, etymology study, and establishment of first attestations. Also, any other annotations or information that has been added to the New Testament (the pragmatic word order, discourse particles, pronominal reference or background events) may be imported into the Romanian Oldest New Testament.

The texts are previously annotated in the UAIC conventions and a program made the automatic transformation (supervised) in the UD conventions. So, UD_Romanian-Nonstandard is a part of the UAIC-RoDia Dependency Treebank (RoDia – for Romanian Diachronic), which is recognized in the international resource catalogue, with the id ISLRN 156-635-615-024-0. At the moment, we included in the UD_Romanian-Nonstandard treebank the 11 documents in which the New Testament (Alba Iulia, 1648) was processed in XML in the UAIC format. These documents have been transformed in the UD conventions and validated.

As seen in this chapter, the process of automatic matching should be supervised by knowledgeable language specialists. But the effort is justified by the great benefits of studying the processes used by the Romanian translators in writing the seventeenth-century holly books.

In the **final chapter** we present conclusions and further work. Natural language processing is an essential activity for the modern world, for the access to the cultural heritage, in which we can perform information searches, automatic translations, summaries.

The specific character of our treebank is that it contains a variety of styles of Romanian, old contemporary and regional, with the intention of being a training corpus for both standard and non-standard Romanian.

We will continue to focus on the study of the Old Romanian, which has four centuries of evolution, and their weight should be balanced, as well as the weight of the styles; for the time being, the church style is predominant in disfavour of the narrative or the legislative one.

The total number of sentences should reach 70,000, in order to allow the proper training of the POS-Tagger on our own corpus and to extract adequate statistics from it. Then, the training of the parser could be conducted on sub-corpora, i.e., centuries and styles, each having around 10,000 sentences.

The accuracy of the syntactic parser on the basic format must be increased, by increasing the training corpus and its consistency. After reaching an accuracy of more than 85% (=LAS), the supervision of the results will be easier and by the bootstrapping method, the training corpus will be increased and well structured. The last tests showed an accuracy increase of LAS=87% on certain sub-corpora, which is encouraging.

The data entry in the POS-tagger lexicon should be permanently continued. The program that extracts new word-lemma-MSD combinations from the corrected treebank still leaves behind numerous errors that need to be removed from the treebank (not from the output of the program; after eliminating the errors and a re-training of the parser, it is hoped that the same errors will not appear again). Through these corrections we ensure both the consistency of the corpus and the preparation of the new data to be introduced in the lexicon (the correct ones remaining).

Another necessary operation would be to separate the chain of automatic processing, so that the output of the POS-tagger can be corrected before the syntactic parser is applied. These programs are currently working in pipeline. The syntactic parser must establish links between the morphological analysis of the parsed word and of its head, in order to propose a direction of subordination and a type of relationship, but such links are sometimes performed to the wrong morphological category and therefore parsing will be mistaken.

The statistical syntactic Malt parser must be trained separately on the two syntactic conventions of annotation. It is always trained on a CONLLU format of sentences, i.e., the UAIC RoDia Dependency Treebank must first be converted in CONLLU, for the training, and after the training, it must be used for the automatic annotation of new texts. The output will be also in CONLLU and another converter will transpose the new automatically annotated text into the XML format of our treebank.

Such converters are necessary for all sorts of applications. The search programs of some sites, as CoRoLa, cannot made searches in another format than XML. The TREEOPS program rule based which can transform the basic format of our treebank into the UD format has rules for the XML conversion.

Regarding the semantic annotation convention, after TREEOPS would do the partial transformation from the basic syntactic convention, to the semantic convention, we could train a statistical parser on the result of the transformation, so that the versatility of the syntactic relations is solved statistically. The question is how many sentences the training gold corpus should have, relative to the total number of 96 semantic dependency relationships, given that the number of 10,000 sentences is considered suitable for a set of approximately 40 tags.

As the large dimensions entail the need for an increasing number of possible options, the training corpus will have an increase more than proportional to the set of tags. The training corpus already transformed, then supervised and completed with semantic information, of 5,566 sentences, is clearly insufficient.

Another way to solve the semantic annotation in the future would be to use the Pattern Dictionary of Romanian Verbs, in which the words to be correlated with the PWN (Princeton WordNet), so that the computer can extract the circumstantial values of the nominal determiners, for example.

The annotation in the semantic format of the treebank will have to be followed and diversified according to the language styles, going from the church to the narrative, legislative and folkloric. With a small number of examples, the results cannot be conclusive.

When there will be at least 15,000 sentences, the sorting mechanism of similar semantic structures could work and we could see what the percentage of occurrences for each type of judgment is represented within the occurrences of a verb in the semantic corpus.

The semantic pattern sorting program should calculate the percentages in which their occurrence is expected in our semantic corpus, as it calculates the percentages in which syntactic patterns are expected in our nonstandard syntactic corpus, and these statistics should serve for the statistical parser orientation.

The next step would be to enter in the treebank several sentences of the type of the wrong one automatically annotated, to analyse them consistently, to insert them in the training gold corpus so that the parser can correctly extract the syntactic annotation rule.

An attribute that we must not abandon or neglect is also the improvement of the tools for the OCR of Old Romanian Cyrillic letters. Creating an Optical Character Recognizer capable to recognize and transform the text into Old Cyrillic editable letters is a big step in the direction of the digital access to the textual cultural heritage, which cannot be achieved only by scanning old books.

Reading the manuscripts would be a further step in retrieving information from old documents. It is difficult because the manuscripts have a polyglot appearance, containing data in Romanian and Slavonic or in Romanian and Medieval Greek.

This activity is based on the compilation of a large inventory of letters in the text, together with their reading, as well as on the increase of the dimensions of the included lexicon, which must nevertheless be prepared for inclusion in the morphological analyser.

Storing texts corrected by linguists and entering them into a database with which the OCR program is trained is another important step. As all these programs, like many others, are based on the corpus, we will continue to involve ourselves in its and their development.

The thesis also includes a bibliography of titles mentioned in the text (with 248 titles), a list of personal publications (with 35 titles, out of which 5 have the grade B, and some the grade D, totalising more than 22 CNATDCU points) and 3 annexes. Following the list of personal publications, we give only the titles referred in this summary.

Original contributions

This thesis describes our work performed during the 7 years of doctoral research (2015-2022), but which is rooted on contributions and publications also developed beyond the specified period. The core of this work resides in the development of a computerised linguistic resource which is essential for the upgrade of the processing machinery of the Romanian language to the level of other important languages in the world: an annotated corpus of syntactic constructions. As such, our main contributions are as follows:

Theoretical contributions:

- a thorough analysis of similar resources developed for Romanian and for other languages;
- the crystallisation of an empirical proof based on examples that the lexical level of the language used in

the Republic of Moldova and in Romania is the same, with minor differences at the syntactic level due to even more freedom in the word order;

- a clear identification of the differences between the UAIC conventions of treebank annotation (best representing the idiosyncrasies of Romanian) and the UD conventions;
- a proposal for augmenting the Romanian treebank with a semantic layer; 14 different circumstantial values are proposed, closely related to their syntactic equivalents;
- proposals for a research methodology to be followed in order to achieve a mature level of development of training and evaluation resources for the Romanian language, covering the whole diversity of registers, styles, diachronicity and synchronicity.

Linguistic resources:

- the Romanian Diachronic Dependency treebank for Romanian – UAIC-RoDia, 34,794 sentences, and 714,377 tokens, available in 3 formats;
- the UD_Romanian-Nonstandard treebank, with nearly 16,000 sentences from a variety of styles of old, contemporary and regional Romanian, a treebank aligned with 87 other treebanks in the big family of Universal Dependencies;

Instruments to process language:

- the UAIC to UD transformer;
- the TREEOPS XML-transformer program (in cooperation) – a tool that transforms semantically non-ambiguous syntactic relationships into semantic relationships;

- the XML-CoNLLU for UDV2 converter (in cooperation);
- a tool that transforms semantically unambiguous syntactic relations into semantic relations;
- a rhyme statistics program (in cooperation);
- PDRoV-RoDia site - a portal for editing, searching and viewing syntactic and semantic dependency structures, together with the repository that hosts the XML + CONLLU structure.

References

1. Barbu-Mititelu, V.: 2013 *Derivational Semantic Network for Romanian*, Bucharest. National Museum of Romanian Literature Press.
2. Bobicev, Victoria, T. Bumbu, V. Lazu, V. Maxim, D. Istrati 2016 Folk poetry for computers: Moldovan Codri’s ballads parsing. Proceedings of the 12th International Conference “Linguistic Resources and Tools for Processing the Romanian Language, p. 39–50.
3. Colhon Mihaela and Radu Simionescu. 2012. Deriving a Statistical Parsing from a Treebank. In Proceedings of International Conference on Web Intelligence, Mining and Semantics WIMS (WIMS). ACM Publishing, Craiova Romania, 34:1–34:8.
4. Cristea, D.: Formalisms and Tools for Description and Natural Language Processing, Iasi, Alexandru Ioan Cuza University Press. (2002).

5. Curteanu Neculai and Alexandru Moruz. 2012. A Procedural DTD Project for Dictionary Entry Parsing Described with Parameterized Grammars. In Proceedings of COGALEX-III – The Third Workshop on Cognitive Aspects of the Lexicon.Bombay, India, 127–136.
6. Diac, Paul, Cătălina Mărănduc, and Mihaela Colhon, Relationships and Sentiment Analysis of Fictional or Real Characters. 2018. In Proceedings of 19th International Conference on Computational Linguistics and Intelligent Text Processing, 18-24 March 2018, Hanoi, Vietnam.
7. Gîfu. D, The Analysis of Diachronic Variation in Romanian Print Press. Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems, NESSUS PhD (2016), pp. 49--53.
8. Hall, J., and, Nilsson, J., 2007. *CoNLL-X Shared Task: Multi-lingual Dependency Parsing,* MSI report 06060, Växjö University, School of Mathematics and Systems Engineering.
9. Ion, R., Ştefãnescu, D.: Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization, in Z. Vetulani, *LTC 2009*. pp. 435-443. Heidelberg, Springer, (2011)
10. Irimia, E,, and Barbu-Mititelu, V., *Two resources developed in the project Semantics Driven Syntactic Parser for Romanian,* in Proceedings of ConsILR, Mălini, 27-29 Oct. 2016, p. 69-78. (2016)
11. Mărănduc, Cătălina, Cenel-Augusto Perez, A Resource for the Written Romanian: the UAIC

Dependency Treebank, in Proceedings of ConsILR, Mălini, 27-29 Oct. pp. 79-90. (2016)

12. Moruz, Al., Dezvoltarea unui adnotator FDG (Functional Dependency Grammar) pentru limba română, lucrare de disertaţie, Facultatea de Informatică, Universitatea „Alexandru Ioan Cuza" Iaşi. (2008)
13. Nivre. Joakim 2003. An efficient algorithm for projective dependency parsing. In Proceedings of the 8th InternationalWorkshop on Parsing Technologies (IWPT). ACL Publishing, Nancy France, 149–160.
14. Trandabăţ, D.: Natural Language Processing Using Semantic Frames, PHD Thesis, Computer Science Faculty, Alexandru Ioan Cuza University of Iasi, Romania. (2010)
15. Tufiș Dan și Dragomirescu Liviu Tiered Tagging Revisited In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004, ELRA, Paris, p. 39-42. (2004)