A Diachronic Corpus for Romanian (RoDia)

Cătălina Mărănduc[1,3], Cenel-Augusto Perez[1], Ludmila Malahov[2], AlexandruColesnicov[2]
[1]Faculty of Computer Science, Al. I. Cuza University, Iași, Romania
E-mail:{catalina.maranduc, augusto.perez}@info.uaic.ro
[2]Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
E-mail: lmalahov@gmail.com, acolesnicov@gmx.com
[3]Academic Institute of Linguistics "IorguIordan – Al. Rosetti" Bucharest, Romania

Abstract

This paper discusses the evolution of a Romanian corpus of the type Dependency Treebank, built at the Al. I. Cuza University of Iasi. The corpus has rich annotations and a balanced structure. The researchers who have built this corpus are interested not only in contemporary and standardized Romanian. A non-standardized aspect, namely, Social Media communication, has been studied. The study of other non-standardized aspects of the language, the regional style and old Romanian, began by creating some sub-corpora, which were still too small. Having the intention to participate at the PROIEL project, that align the oldest New Testaments written in Latin, Greek, Slavonic and Armenian languages, we choose to annotate the first printed Romanian New Testament from Alba Iulia (1648). The book, having approximately 250,000 tokens and 12,000 sentences, will be entirely annotated in several formats. We began by applying over the previous, morphological processing with classical syntactic annotation of the first quarter, the automated one. . We applied these tools to a fragment from the second edition in the modern Latin script. The existence of books printed in the old Cyrillic alphabet before the middle of the nineteenth century is a common problem for Romania and The Republic of Moldova, countries where the Romanian language is spoken and a problem to be solved by the joint efforts of NLP researchers in the two countries. A first fragment of the Alba Iulia New Testament (1648) was transformed into an editable Cyrillic text by an OCR program and then transposed to the Latin alphabet by the researchers of the Institute of Mathematics and Computer Science of Chisinau, which was the first such operation in practice. The editable text in the Cyrillic script obtained in the first step of the process was checked by comparing it with the printed old book, and then was wrapped in the XML format, obtaining the second form of the first 500 annotated and manually supervised sentences. The Latin form obtained in the second step of the processing was compared with the second edition of the book, written by priests, without the intention to actualize, or normalize the text. Our purpose has been to use the entirely annotated and checked book for the extraction of an old lexicon to be introduced in the tools and also for the training of these tools on Old Romanian. Simultaneously, the work of the regional sub-corpus has begun, with the intention to include, in future, the south Danube dialects of Romanian. The transformation of the syntactic classic treebank in the UD format and in another semantic format has also begun. This old Romanian treebank will also be a balanced treebank, including not only religious books, but also historical texts (chronicles), legal texts (codices), and anonymous folk tales.

1. Introduction

The heritage of information contained in the old texts needs to be processed, in order to applying the programs for information retrieval and questions answering, for machine translations or for automatically texts resuming, operations in whose absence this information is not accessible to contemporary readers or researchers. All these operations are based on the POS-tagging or on the syntactic parsing; the texts must be previously annotated on morphological and syntactic information.

The NLP group of the Faculty of Computer Science of Iasi has built a balanced corpus of the type Dependency treebank, called UAIC-RoDepTb. At present, it has 11,613 sentences and over 213,000 automatically annotated and manually supervised tokens. The treebank contains standardized and non-standardized language (2,500 Social Media sentences). The training of the UAIC POS-tagger **Fehler! Verweisquelle konnte nicht gefunden werden.** and of the syntactic parser on Social Media was a difficult task. The corpus contains more styles of the contemporary Romanian, such as Folklore, journals, legal style, Romanian and foreign fiction, or the popularization scientific style (Wikipedia). This treebank contains 1,000 sentences in Old Romanian, the nucleus of the RoDia (Romanian Diachronic) corpus.

Old Romanian is also a non-standardized variant of the language, because the rules have not yet been established and each writer applied his own principles to transcribe the spoken language or to change the form of inflected words. The syntactic peculiarities are also very diverse. Our tools have not yet been trained on this type of texts.

Building this corpus, we have also built or trained some tools for Romanian language processing that have a good accuracy percentage for contemporary standardized Romanian.

The academic Institutes of Linguistics in Iasi and Bucharest have big collections of non-annotated texts in all the styles and from all the periods of the Romanian language evolution. The books are printed and processed by optical character recognized programs. The (non-supervised) texts obtained, have more errors in the case of old texts.

There are a lot of books that could not yet be processed by the OCR programs: the Romanian texts written in Old Cyrillic alphabets. The researchers of Romania and of The Republic of Moldova have the same problem. The two countries constituted a single state in the past, the historical documents (written in Old Cyrillic) are common, and the regional variants of Romanian spoken in the two countries, with few differences, are mutual understandable.

We have decided to continue together increasing our balanced treebank. Our purpose is to build a big training gold corpus for old Romanian, to collect the data from it and also to build variants of our tools that can automatically process the old text with good accuracy.

2. Directions for Increasing the Corpus

The balanced corpus will be increased by adding a lot of regional variants. For the moment, the introduction of popular regional texts collected in the two countries is in progress, adding the folklore in verse from all the regions. That is also a non-standard style, because each regional variant has its own regularities, and we must train our tools on each of them. Comparisons between the regional peculiarities and statistical studies of differences will become possible. For these comparisons, the sub-treebank of The Republic of Moldova must include other communication styles, for example journalese or fiction.

In future, we intend to add also, some texts from the South Danube dialects of Romanian. Aromanian **Fehler! Verweisquelle konnte nicht gefunden werden.**, Istro-Romanian **Fehler! Verweisquelle konnte nicht gefunden werden.**, and Megleno-Romanian **Fehler! Verweisquelle konnte nicht gefunden werden.**, **Fehler! Verweisquelle konnte nicht gefunden werden.** are very different from the language spoken in our two countries and are not mutually intelligible. It will be a very difficult task to process them. We began the collection of texts (with very different conventions of phonetic transcription!) but we must build other tools for processing them. It is a task of great importance, because the dialects are endangered languages, especially Istro-Romanian and Megleno-Romanian, which are spoken only in families and have not developed a written aspect.

The annotation of old Romanian is our priority. We chose to begin with the New Testament of Alba Iulia (1648), the first printed New Testament in Romanian, with the intention of affiliating it to the PROIEL project **Fehler! Verweisquelle konnte nicht gefunden werden.**, which aims to align the oldest Latin, Greek, Slavonic and Armenian NTs. After the training of tools on this gold corpus, the balanced treebank must be completed with other styles of the old language, with the historical chronicles style, with legalese (codices of laws and documents), with anonymous folk tales, and even a cookbook.

The annotation conventions specific to our treebank will be kept in order to create a very big corpus without inconsistencies, that being the condition for increasing the accuracy of tools on all the language variants. But a program for their transformation in the international conventions of the treebanks reunited in the Universal Dependencies (UD) project is also needed. Another program will transform the classical annotation into a semantic one. The construction and the training of these programs have begun.

3. Related Work

As regards diachronic corpora, we can cite two recent conferences: "International Conference on Practical Applications of Language" PALC 23-24 October 2015, Lodz[1], and "Diachronic Corpora, Genre, and Language Change", 8-9 April 2016, Nottingham[2].

The pragmatic interpretation of ancient texts is a new direction of research. In the Proceedings of the 12th International Pragmatics Conference in Manchester in 2011 there are chapters based on diachronic pragmatic interpretation of several corpora **Fehler! Verweisquelle konnte nicht gefunden werden.**. In 0, the alignment of the oldest New Testament also has a pragmatic purpose.

The diachronic corpus of Italian is described in 0.

There exist also corpora for the dead languages: "The Diachronic Corpus of Sumerian Literature" (DCSL) a web-based corpus of Sumerian literature and concerning the entire history of the Mesopotamian civilization[3].

There are several texts that describe the creation of diachronic corpora, their purpose, their difficulties, and their usability. A historical American English corpus is presented in 0; a similar corpus for old French is described in 0; finally, the creation of a German diachronic corpus is described in 0. A comparison between the historical Spanish and Portuguese corpora is made in 0.

In 0, a computational morphological description of old Swedish is made, and the 0 paper is a guideline for the creation of a richly annotated corpus. The paper 0 is another guideline, intended for linguists who are uninformed about the perspectives created for their research by linguistic of corpora.

4. The OCR Program for the Old Cyrillic Books

The Romanian linguists published a big number of editions with comments for each important book written with old Cyrillic texts, including their transcription in the Latin alphabet and final indexes or glossaries. But these philological editions are not usable for our purpose. They contain controvertible theories of the sixteenth and seventeenth century's pronunciation and interpretative transcription. The indexes at the end of these books would have been very useful, but if we introduced them in the lexicon of processing tools, they

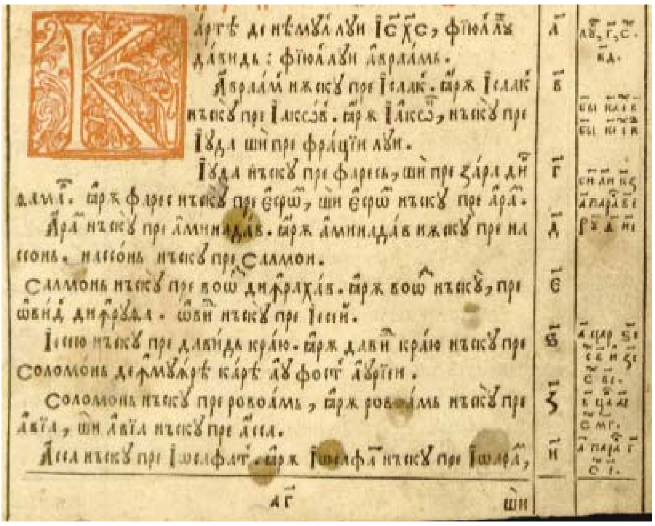
[1] http://palc.uni.lodz.pl/

[2] https://www.nottingham.ac.uk/conference/fac-arts/clas/dcglc/home.aspx/

[3] http://dcsl.orinst.ox.ac.uk/

would not find anything in the texts, because the form of words was an interpretative spelling, or was replaced with the lemma (the form find in the contemporary dictionaries).

The first printed Romanian New Testament, chosen for beginning the gold corpus for old Romanian, is relatively well preserved and written in the old Cyrillic alphabet. See Figure 1.



Кáртѣ де нѣмꙋлꙋи Іс Хс, фїюлꙋи
Дáвида: фїюлꙋи Аврáмь. а
Аврáм нъскꙋ пре Ісаак. іар Ісаак
нъскꙋ пре Іаковь. іар Іаковь, нъскꙋ пре в
Іꙋда ши пре фрáцїи лꙋи.
Іꙋда нъскꙋ пре Фаресь, ши пре Зáра дин г
Тамі. іар Фарес нъскꙋ пре Есрѡм, ши Есрѡм нъскꙋ пре Арам.
Арам нъскꙋ пре Аминадав. іар Аминадав нъскꙋ пре На д
ссонь. Нассóнь нъскꙋ пре Салмон.
Салмонь нъскꙋ пре Воѡз дин Рахав. іар Воѡз нъскꙋ, пре е
Ѡвид дин Рꙋфа. Ѡвид нъскꙋ пре Іесей.
Іессеѡ нъскꙋ пре Давида крáю. іар Давид крáю нъскꙋ пре ѕ
Соломонь дин мꙋіаре кáре аꙋ фост а Ꙋрїеи.
Соломонь нъскꙋ пре Ровоáмь, іар Ровоáмь нъскꙋ пре з
Авїа, ши Авїа нъскꙋ пре Аса.
Аса нъскꙋ пре Іѡсафат. іар Іѡсафат нъскꙋ пре Іѡрам, и

аг ши

Fig.: 1. A fragment of the beginning of Alba Iulia New Testament (1648).

Fortunately, the second edition of the Alba Iulia New Testament was published by priests, who did not practice interpretive transcription. We have processed its first 3,000 sentences with our contemporary Romanian tools. The result, with a modest accuracy, is being carefully checked manually and will be used to train the tools on old Romanian. From a philological point of view, this does not mean that we annotate the first New Testament printed in Romanian. The books in the PROIEL project are written with their original Greek, Slavonic or Armenian characters.

Therefore, researchers at the Institute of Mathematics and Computer Science in Chisinau have performed by their OCR program, a version of a first portion of the text. It is the first time that a Cyrillic Romanian text of the seventeenth century is obtained by an OCR program.

This OCR program has been trained on Romanian texts printed after 1945 in The Republic of Moldova with the modern Cyrillic alphabet. In Romania, books were printed with Latin characters starting from the middle of the nineteenth century. The program has been trained also on texts printed in Romania at the beginning of the nineteenth century, with a good accuracy, which we cannot hope to achieve for texts of the 17th century.

The researchers of Chisinau have processed the print in two steps. In the first step, they obtained an editable form of the book with old Cyrillic characters. The computational linguists of Iasi have remarked that two characters are not in the ”times_extro” set of letters and are not recognized by the other formats of documents, except word documents, because the set of characters are not all in the ASCHI set.

The computational linguists of Iasi replaced the sign ꙟ with an arrow ”↑”, and they maintained щ in the form that has been used by the researchers of Chisinau, ”щ”, although this letter is spelled șc, and not șt, as it should be. For the letter ȣ they found a similar character, ”ȣ”, and, with this set of letters, they succeeded in replacing the Latin characters in the checked xml. A resulting sentence, annotated in the UD conventions, is shown in Figure 2.

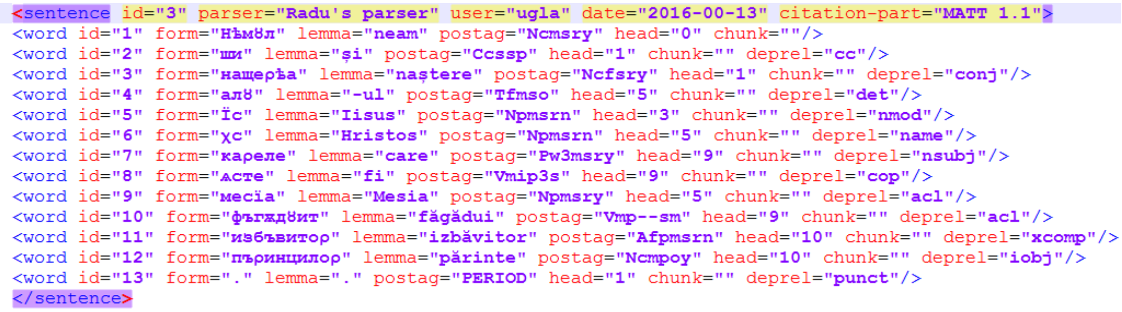
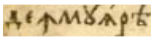
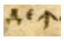
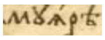

```
<sentence id="3" parser="Radu's parser" user="ugla" date="2016-00-13" citation-part="MATT 1.1">
<word id="1" form="Нѣмȣл" lemma="neam" postag="Ncmsry" head="0" chunk=""/>
<word id="2" form="ши" lemma="și" postag="Ccssp" head="1" chunk="" deprel="cc"/>
<word id="3" form="нащерѣа" lemma="naștere" postag="Ncfsry" head="1" chunk="" deprel="conj"/>
<word id="4" form="алȣ" lemma="-ul" postag="Tfmso" head="5" chunk="" deprel="det"/>
<word id="5" form="Їс" lemma="Iisus" postag="Npmsrn" head="3" chunk="" deprel="nmod"/>
<word id="6" form="χс" lemma="Hristos" postag="Npmsrn" head="5" chunk="" deprel="name"/>
<word id="7" form="кареле" lemma="care" postag="Pw3msry" head="9" chunk="" deprel="nsubj"/>
<word id="8" form="ѧсте" lemma="fi" postag="Vmip3s" head="9" chunk="" deprel="cop"/>
<word id="9" form="месїа" lemma="Mesia" postag="Npmsry" head="5" chunk="" deprel="acl"/>
<word id="10" form="фъгѫдȣит" lemma="făgădui" postag="Vmp--sm" head="9" chunk="" deprel="acl"/>
<word id="11" form="избъвитор" lemma="izbăvitor" postag="Afpmsrn" head="10" chunk="" deprel="xcomp"/>
<word id="12" form="пъринцилор" lemma="părinte" postag="Ncmpoy" head="10" chunk="" deprel="iobj"/>
<word id="13" form="." lemma="." postag="PERIOD" head="1" chunk="" deprel="punct"/>
</sentence>
```

Fig. 2: The first sentence with old Cyrillic, annotated in UD conventions (Conception and Birth of Jesus Christ, which is the promised Messiah, the parents’ deliverer.).

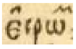
Simultaneously with the replacement of the word form in Latin characters with the word form in Cyrillic characters, the linguists confronted the transcription with the printed book and they signalled the errors of the Chisinau researchers.

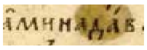
a) A big problem was the numbers of chapters and paragraphs to be preserved. These numbers are universal for any version of the Christian New Testament and essential for the alignment of our New Testament with the other books in PROIEL. We decided to introduce them in the general data of each sentence. In the PROIEL project, they were repeated for each word, because in the CONLLU format there is no information about sentences. Our NLP group has built a program for the XML to CONLLU format transposition; there will be added a function to expand the ”citation part” (see fig. 2, first line) of each word.

b) Therefore, in the printed book, these numbers are in a separated column and are written with old Cyrillic letters. The sign а҃ represents the first paragraph, в҃ the second, г҃ the third, and so on. (See Figure 1). The list of numbers-letters and the mark by an upper score will be introduced in the OCR program. These numbers-letters appear only in the oldest books.

c) The OCR must include a splitter that can cross out with spaces the words where they are not cut in the originally printed text, using a big lexicon for old Romanian. For example, in Figure 1,  is a string of signs that can be separated,  being a preposition, ”deîn, den”, ”*from*” and , muiarea< Lat. mulier.”*the woman*”. This peculiarity is specific of the first printed books, probably the ”white” letter did not yet exist and the letters were placed at various distances that might chance to disappear at the moment when they were pressed.

d) The book contains a lot of abbreviated words. The researchers of Chisinau made a list of the individual abbreviations and they wanted the program to replace the abbreviations with the entire words. Perhaps this was not compulsory, since abbreviations exist in contemporary languages, too; it is enough to put the list of abbreviations at the beginning of the document.

e) Frequently, there are no abbreviated words, but words with some overwritten letters. For example, the word  (Esrom, proper noun) has an accent and the letter m placed over the letter o. This peculiarity will be a permanent source for decreasing the accuracy percentage and these words will be manually checked.

f) Some of the missing letters or words were due to stains. This difficulty could be overcome by giving up the Scan Taylor program (which turned the text black) and by increasing the resolution. The word  (Aminadav) wasn’t entirely recognized.

5. Transcription Difficulties

The second step of the OCR processing created at Chisinau is the transliteration of old Cyrillic letters as Latin characters. The researchers of Chisinau have a POS-tagger that can annotate with morphological information both Romanian texts written with Latin letters and Romanian ones written with modern Cyrillic characters; the program was adopted for old Cyrillic (introducing new characters) and for the operation of transposition (introducing new rules).

The computational linguists of Iasi compared the text with Latin letters obtained by the researchers of Chisinau with the second edition of the New Testament (used in the automatic morphological and syntactic annotation of the first 3000 sentences) with Latin characters, obtained by the Abby Fine Reader.12 program. A fragment of the comparison is shown in Table 1, most of it corresponding to the image of Figure 1.

Table 1.Comparison between the second edition and the OCR of the first edition.

The New Testament at Alba Iulia (1648) Second Edition, with Latin Letters, Alba Iulia, Publisher Romanian Orthodox Diocese of Alba Iulia, 1988.	**The New Testament at Alba Iulia (1648) First Printed in Romanian by Simion Ştefan, Metropolitan Bishop of Transylvania. Ocreized at IMATH Chișinău 02.08.2016.**
Neamul şi naşterea lui Iisus Hristos carele iaste Mesia făgăduit izbăvitor părinţilor. 1. Cartea de neamul lui Iisus Hristos, fiiul lui David, fiiul lui Avraam. 2. Avraam născu pre Isaac, iară Isaac născu pre Iacov, iară Iacov născu pre Iuda şi pre fraţii lui. 3. Iuda născu pre Fares şi pre Zara din Tamar, iară Fares născu pre Esrom şi Esrom născu pre Aram. 4. Aram născu pre Aminadav, iară Aminadav născu pre Nasson, Nasson născu pre Salmon. 5. Salmon născu pre Vooz din Rahav, iară Vooz născu pre Ovid din Ruta, Ovid născu pre Isei. 6. Ieseiu născu pre David craiu, iară David craiu născu pre Solomon, den muiarea carea au fost Uriei. 7. Solomon născu pre Rovoam, iară Rovoam născu pre Avia şi Avia născu pre Assa. 8. Assa născu pre Ioasafat, iară Ioasafat născu pre Ioaram şi Ioaram născu pre Oziia. 9. Oziia născu pre Ioatam, iară Ioatam născu pre Ahaz, iară Ahaz născu pre Ezechiia. 10. Ezechiia născu pre Manasia, iară Manasiia născu pre Amon. 11. Amon născu pre Iosian, iară Iosian născu pre Ehonian şi pre	N~~H~~eamul şi nașterea a lui is hs,carele iaste mesia făgăduit izbăvitor părinților. 1. Cartea de neamul lui IS hS, fiiul lui david : fiiul lui avraam. 2. Avraam născu pre Isaac. iară Isaac născu pre Iacov. iară Iacov,născu pre Iuda şi pre frații lui . 3. Iuda născu pre fares, şi pre zara din Thamar. iară fares născu pre Esrom, şi Esrom născu pre aram. 4. Aram născu pre aminadav . iară aminadav născu pre nasson. nasson născu pre Salmon. 5. Salmon născupre vooz diî rahav. iară vooz născu, pre ovid diîrutha . ovid născupre Iesei. 6. Ieseiu născu pre david craiu. iară david craiu născu pre Solomon deîn muiarea carea au fost a uriei . 7. Solomon născu pre rovoam , iară rovoam născu pre avia, şi avia născu pre assa. 8. Assa născu pre Iosafat. şi Iosafat născu pre Ioaram. şi Ioaram născu pre ozia. 9. Ozia născu pre Ioaɵam. iară Ioaɵam născu pre ahaz. iară ahaz, născu pre Ezechia. 10. Ezechia născu pre manasia. iară manasia născu pre amon. 11. Amon născu pre Iosian, iară Iosiae născu pre Ehonian şi pre frații lui, în vavvlon.

The red characters are missing in the OCR version, and the text with yellow shading is missing in the second edition form that is used in the automatic annotation.

The formal observations are that the letter џ does not turn into the Latin j, but into ge, gi. The cyrillic letter ”s” is traditionally spelled dz, the letter ɵ is spelled th, and the group ”ou” is always spelled ”u”.

Examples:↑џер8л = îngerul, not înjerul ”the angel”; д8мнеse8=dumnedzeu, and not dumneseu ”*God*”, оунде=unde, not ounde ”*where*”. Other more difficult issues were successfully settled. For example, the letter ѧ and ѩ is spelled ”ia” and їѧ must be spelled ”iia”;ю is spelled ”iu” therefore їю is spelled ”iiu”; the letter ↑ can be spelled ”î” ”îm” or ”în”, for ”ă” the old text sometimes uses ѫ, sometimes, ъ, and so on.

The first theoretical observation to be made is that our purpose is to study the peculiarities of the old text, but not to remove them by ”correcting” the text, i.e. by applying rules/norms which did not exist in the studied period of the language evolution.

The bishop Simion Ștefan was a cultivated man, he was aware of the European ideas of his time, demanding that the religious service and the sacred books be in the language of the people, therefore he wrote without capital letters because in this period there was no rule for the capitalization of proper nouns and of the pronouns co-referential with the noun of the Divinity.

Spelling correction is made for the use of a wide audience of non-specialists who are not interested in old books and will probably not read them. On the contrary, the persons interested in the real appearance of the ancient text will not have access to it.

We believe we must enter in the lexicon of processing tools all the forms existing in the old language, in order for them to be recognized and correctly analysed when found in other texts. For example, if we introduce in the lexicon only capitalized proper names, the ones encountered without capitalization will not be correctly annotated.

It is also worth noting that, at that time, the orthographical convention of words joined by a hyphen did not exist. For example, *aflăsă* means *află-se*, *se află* “*there was*”. Therefore, we must annotate two words because the string has the meaning of two words. The inversion of word marks, not accepted in the contemporary language, is frequently used in ancient texts. Being frequent, the statistical tools will probably be easily trained on it in this respect.

6. Conclusions and Future Work

After taking the steps consisting in checking the annotation and the OCR obtained by confronting it with the printed book, we can say that we have actually annotated the first printed Romanian New Testament. It is a good idea to compare two transcriptions, (both without the intention to actualize, interpret or correct them), because the text is very ancient and difficult. Each of the versions compared has mistaken that the other does not have, and problems better solved than the other. The shading parts (yellow in the table) are better solved by the version of Chisinau.

For the second section of 3,000 sentences, we intend to introduce the Chisinau version in the programs for the automatic morphological and syntactic annotation that will be manually checked. Therefore, the comparison of the second edition must continue. After each step, the programs will be ameliorated by the introduction of the correct information in the lexicon and by the training on an increased gold corpus. By the training with the Cyrillic variant, the Iasi POS-tagger will be able to annotate Romanian ancient texts written in the Cyrillic and Latin alphabets. Perhaps some Romanian linguists will want to study the Cyrillic version of texts, without transposing them into the modern alphabet, as specialists in old Slavic or in old Greek do.

The treebank has three layers (conventions of annotation). We first annotate each text in the classical syntactic convention that our tools are trained in and that contains a big quantity of information. It can be automatically transposed in the UD convention, which is the international link between more than 30 treebanks, but has less information. Also the classical syntactic annotation can be automatically transposed in the semantic treebank. The transformations will be supervised and the programs for the transformations will be trained on the checked versions.

After the full processing and supervision, the New Testament will be transformed in the UD convention and in the CONLLU format and affiliated to the PROEL project, that will align it with the other old New Testament versions and will study the pragmatic connections between the sentences and other pragmatic peculiarities. The New Testament will be submitted in four forms: in XML and in CONLLU, with Latin and with Cyrillic characters.

Meanwhile, another team will continue increasing the treebank with popular regional texts from all the regions of our two countries. The tools will be trained on each of these non-standardized variants of the language.

7. References

Number	Title, author(s)	Year
Fehler! Verweisquelle konnte nicht gefunden werden.	**The Current State of Megleno-Romanians. Megleno-Romanian, an Endangered idiom.** P. Atanasov Memoria Ethnologica, XIV, 52-53, July-December, p. 30–37.	2014
Fehler! Verweisquelle konnte nicht gefunden werden.	**Something Old, Something New: A Computational Morphological Description of Old Swedish.** L. Borin, M. Forsberg. LREC 2008 Workshop on Language Technology for Cultural Heritage Data pp. 9–16.	2008
Fehler! Verweisquelle konnte nicht gefunden werden.	**Aromanians and Aromanian Dialect in Contemporary Consciousness.** M. Caragiu-Marioțeanu Romanian Academy Publisher, Bucharest.	2006.

Fehler! Verweisquelle konnte nicht gefunden werden.	**Creating Useful Historical Corpora: A Comparison of CORDE, the Corpus del Español, and the Corpus do Português.** M. Davies *Diacronía de las Lenguas Iberorromances: Nuevas Perspectivas desde la Lingüística de Corpus*, ed. Andrés Enrique-Arias. Frankfurt/Madrid: Vervuert/Iberoamericana, p. 137-166.	2010
Fehler! Verweisquelle konnte nicht gefunden werden.	**Expanding Horizons in Historical Linguistics with the 400 Million Word Corpus of Historical American English**. M. Davies Corpora 7, p. 121-157.	2012
Fehler! Verweisquelle konnte nicht gefunden werden.	**Challenges in Modelling a Richly Annotated Diachronic Corpus of German.** S. Dipper, L. Faulstich, U. Leser, A. Ludeling, Proceedings of LREC	2010
Fehler! Verweisquelle konnte nicht gefunden werden.	**Il ComuneIstro-romeno di Valdarsa.** N. Feresini, Edizioni Italo Svevo. Trieste.	1996
Fehler! Verweisquelle konnte nicht gefunden werden.	**Creating a Parallel Treebank of the Old Indo-European Bible Translations** D. T. T. Haug, M. L. Jøhndal. C. Sporleder, K. Ribarov (eds.). Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data p. 27-34.	2008
Fehler! Verweisquelle konnte nicht gefunden werden.	**The Islamisation of the MeglenVlachs (Megleno-Romanians): The Village of Nânti (Nótia) and the "Nântinets" in Present-Day Turkey,** T. Kahl, Nationalities Papers, 34:01, p. 80-81.	2006
Fehler! Verweisquelle konnte nicht gefunden werden.	**Le Corpus 'Voies du Français' : de l'Elaboration à l'Annotation.** F. Martineau, C. Diaconescu, P. Hirschbühler, Kunstmann, P. & Stein, A. (eds) Le Nouveau Corpus d'Amsterdam. Actes de l'Atelier de Lauterbad, Stuttgart : Steiner, pp. 121-142.	2007
Fehler! Verweisquelle konnte nicht gefunden werden.	**Corpus-based Language Studies: An Advanced Resource Book.** T. McEnery, R. Xiao, Y. Tono Routledge publisher, London, New York, 389 pp.	2006
Fehler! Verweisquelle konnte nicht	**The DiaCORIS project: a diachronic corpus of written Italian**. C. Onelli, D. Proietti, C. Seidenari Proceedings LREC pp. 1212-1215.	2006

gefunden werden.		
Fehler! Verweisquelle konnte nicht gefunden werden.	**Hybrid POS Tagger** R. Simionescu Language Resources and Tools in Industrial Applications, Eurolan summer school.	2011
Fehler! Verweisquelle konnte nicht gefunden werden.	**Diachronic Corpus Pragmatics** A. Taavitsainen, H. Jucker, J. Tuominen (Eds.) John Benjamins publisher, Amsterdam, 335 p.	2014