A Diachronic Corpus for Romanian (RoDia)

Ludmila Malahov
Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova - Chisinau
Republic of Moldova
lmalahov@gmail.com

Cătălina Mărănduc
Faculty of Computer Science, “Al. I. Cuza” University Iasi, Romania
Academic Institute of Linguistics “Iorgu Iordan Al. Rosetti”, Bucharest, Romania
catalinamaranduc@gmail.com

Alexandru Colesnicov
Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova - Chisinau
Republic of Moldova
acolesnicov@gmx.com

catalinamaranduc@gmail.com

Abstract

This paper describes a Romanian Dependency Treebank, built at the Al. I. Cuza University (UAIC), and a special OCR techniques used to build it. The corpus has rich morphological and syntactic annotation. There are few annotated representative corpora in Romanian, and the existent ones are mainly focused on the contemporary Romanian standard. The corpus described below is focused on the non-standard aspects of the language, the Regional and the Old Romanian. Having the intention to participate at the PROIEL project, which aligns oldest New Testaments, we annotate the first printed Romanian New Testament (Alba Iulia, 1648). We began by applying the UAIC tools for the morphological and syntactic processing of Contemporary Romanian over the books first quarter (second edition). By carefully manually correcting the result of the automated annotation (having a modest accuracy) we obtained a sub-corpus for the training of tools for the Old Romanian processing. But the first edition of the New Testament is written in Cyrillic letters. The existence of books printed in the Old Cyrillic alphabet is a common problem for Romania and The Republic of Moldova, countries where the Romanian is spoken; a problem to solve by the joint efforts of the NLP researchers in the two countries.

1 Introduction

The UAIC-RoDia-DepTb is a balanced corpus of the Dependency treebank type, including all the styles of Romanian, and focusing on non-standard ones. The corpus contains more styles of Romanian, such as folklore, journals, FrameNet translated in Romanian, legal texts (Acquis communautaire), Romanian and foreign fiction, social media (chat) popularization scientific style (Wikipedia), quotations (also poetry). The treebank has now 16,187 sentences, with 322,404 tokens, punctuation included, between which 5,723 sentences in Old Romanian are also included (99,749 tokens).

Preserving the cultural heritage means not only scanning the old books, so as to prevent their loss by paper damaging. We also need to read and annotate the information contained in the old texts; it needs to be preprocessed, in order to apply to it the programs for information retrieval, for question answering, for machine translations, or for automatic 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. A POS-tagger for Old Romanian was built, having an accuracy of 91.66% (Mărănduc et al., 2017).

But the first step to enable processing these old texts is to transform their photos into the editable text format. This step is the most difficult one, because old Romanian texts, although the language has Latin origin, are written in Cyrillic letters of all forms and in all orthographic conventions encountered in the neighboring slave peoples.

Old Romanian is also a non-standard 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. Building this corpus, we have also built or trained some tools for Roma-
nian language processing; they have a good accuracy percentage for contemporary standard Romanian. A clone of the POS tagger for processing Old Romanian was built, but the syntactic parser has not yet been sufficiently trained on this type of texts.

A first fragment of the New Testament (1648, first edition) was transformed into an editable Cyrillic text by an OCR program that began operating in August 2016. The printed book will be entirely transposed to editable Cyrillic letters and then to the Latin alphabet by the researchers of the Institute of Mathematics and Computer Science of Chisinau; this is 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 2,500 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, manually transposed in Latin letters 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 (Universal Dependencies) format and in another semantic format has also begun. The first part of the New Testament, the four Gospels, has been entirely annotated and manually corrected in the classic syntactic format and in the new logical-semantic one.

The Old Romanian sub-corpus will also be a balanced treebank, including not only religious books, but also historical texts (chronicles), legal texts (codices), anonymous folk tales, documents ranged by centuries, and even a cookbook.

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 OCR (Optical Character Recognizer) programs. The (non-supervised) texts obtained have more errors in the case of books printed in the 19th century. But books from the sixteenth to the eighteenth centuries, the Romanian texts written in Old Cyrillic alphabets, could not be processed with an accessible OCR program, such as ABBYY Fine Reader 12.

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 minor differences, are mutually 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 a variant of our syntactic parser that can automatically process the old text with good accuracy. A lexicon obtained from this corpus has been included in the POS-tagger for Old Romanian and (with Cyrillic letters) in the OCR program. We continue to increase the lexicon in order to augment the accuracy of these tools. We use the bootstrapping method, and the checked sentences are always added to the gold corpus for the training of tools.

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 regional variants of Romanian. That is also a non-standard style, because each regional variant of the language has its own regularities, and we must train our tools on each of them.

By introducing in the lexicon of the POS tagger all the lexical or spelling variants of Romanian words, extracted from the Thesaurus Dictionary, we introduced in fact the archaic and the regional lexical variants from all over Romania, but we do not know whether there are any variants not included in this dictionary in the Republic of Moldova. 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

\[\text{http://edtlr.info.uaic.ro/}\]
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 (Pragmatic Resources in Old Indo-European Languages)³ (Haug and Johndal, 2008), which aims to align the oldest Latin, Greek, Slavonic and Armenian New Testaments.

After the training of tools on this gold corpus, the balanced treebank must be completed with other styles of the old language.

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. This classic syntactic format is considered the pivot for the transformation into other formats; having a big amount of semantic information, we can develop these new formats and obtain a training corpus for a semantic parser, which will be trained with priority on old Romanian. The transformation into UD (Universal Dependencies)⁴ conventions is also in progress, because in this way we can compare our language with the other 30 languages affiliated, but the transformation is accompanied by the loss of one part of the semantic information.

3 Related Work

As regards diachronic corpora, we can cite three recent conferences: the International Conference on Practical Applications of Language PALT Lodz⁵, 23-24 October 2015, the Diachronic Corpora, Genre and Language Change, Nottingham⁶ 8-9 April 2016, and the conference Digital Access of Textual Cultural Heritage DATECH Göttingen, 1-2 June 2017⁷.

The need for a research infrastructure for the study of historical textual resources by digitization is increasingly recognized by the historical research community. Historical documents are being digitized on a vast scale in cultural heritage and digital library projects in many countries. Modern linguistics studies increasingly pay attention to diachronic and dialectal variations of lan-

3http://www.hf.uio.no/ifikk/english/research/projects/proiel/
4http://universaldependencies.org/
5http://palc.uni.lodz.pl
6https://www.nottingham.ac.uk/conference/facarts/class/dcglc/home.aspx.
7http://ddays.digitisation.eu/datech-2017/

⁸http://dcsl.orinst.ox.ac.uk

guages, studying them with modern formalisms, from the point of view of modern theories. Linguistic studies based on corpora get expanded and grow in importance. Digitized historical corpora are already created for many languages. 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 (Taaavitsainen et al., 2014). In Haug and Johndal (2008), the alignment of the oldest New Testament also has a pragmatic purpose. The diachronic corpus of Italian is described in Onelli et al. (2006). There also exist corpora for the dead languages: DCSL (The Diachronic Corpus of Sumerian Literature), a web-based corpus concerning the entire history of the Mesopotamian civilization⁸. There are several papers that describe the creation of diachronic corpora, their purpose, their difficulties, and their usability. A historical American English corpus is presented in Davies (2012); the paper Yanez-Bouza (2011) describes the history of the creation of ARCHER (A Representative Corpus of Historical English Registers), and a multigenre historical corpus. A similar corpus for Old French is described in Martineau et al. (2007). The paper Stein (2008) is a syntactic annotation study for another Old French corpus. Finally, the creation of a German diachronic corpus is described in Dipper et al. (2010). A comparison between the historical Spanish and Portuguese corpora is made in Davies (2010). In Borin et al. (2010), a computational morphological description of Old Swedish is made, and the paper Dipper et al. (2010) is a guideline for the creation of a richly annotated corpus. The paper Martineau et al. (2007) is another guideline, intended for linguists who are uninformed about the perspectives created for their research by linguistic of corpora. We also consulted an entire book collecting papers interested in the corpus based linguistic methods and studies (McEnery et al., 2006).

BLARK (Basic Language Resource Kit) is a software for the research in the field of social-humanities, i.e. a kit for the automatic annotation of resources in this domain (Borin and Forsberg, 2008). East European languages have several preprocessing issues. For example, some of them used Cyrillic and Latin scripts in various
periods of time; consequently, some documents need to be transliterated before further processing (Gruszczynski and Ogrodniczuk, 2015).

4 The OCR Program for the Old Cyrillic Books

Romanian linguists published a big number of editions with comments for each important book written with Old Cyrillic letters, 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 centuries pronunciation and use an interpretive 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 would not find anything in the texts, because the form of the words was an interpretive spelling, or was replaced with the lemma (the form found in contemporary dictionaries). The first printed Romanian New Testament, chosen for beginning the old Romanian gold corpus, is relatively well preserved. (See Figure 1)

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, whose accuracy is modest, is being carefully manually corrected 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. But we did not find another solution to start the project, breaking the vicious circle - we cannot have processing tools because we do not have a training corpus and we cannot have a corpus because we do not have automatic annotation tools.

Then, the researchers at the Institute of Mathematics and Computer Science in Chisinau have been able to process the first portion of the text, with ABBYY FineReader OCR program (AFR), and a specially developed tool pack of add-ons for AFR. It is the first time that a Cyrillic Romanian text of the seventeenth century has been obtained by an OCR program (Colesnicov et al., 2016). AFR has initially 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 (due to a law given by the Prince Al. I. Cuza).

The OCR for the Romanian written with Old Cyrillic alphabet is based on the AFR features allowing the introduction of new languages, with their alphabets and lexicons (dictionaries). AFR has also been trained on texts printed in Romania at the beginning of the nineteenth century, with sufficient good accuracy, something 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 (Cojocaru et al., 2017). The computational linguists of Iasi have remarked that three characters are not in the “times extro” set. Consequently, they are not recognized by the computer, because the set of characters are not all
Engl. (KJV, Matthew 8): 27. But the men marvelled, saying, What manner of man is this, that even the winds and the sea obey him! 28. And when he was come to the other side into the country of the Gergesenes, there met him two possessed with devils, coming out of the tombs, exceeding fierce, so that no man might pass by that way. 29. And, behold, they cried out, saying, What have we to do with thee, Jesus, thou Son of God? art thou come hither to torment us before the time?

Figure 3: A fragment of the Alba Iulia New Testament (1648)

in the ASCII set. The computational linguists of Iasi replaced the letters with other signs (see Figure 4, first row), and they maintained the letter şc in the form used by the researchers of Chisinau, although it replaces the letter ş, and not şc. With this set of letters, they succeeded in replacing the Latin characters in the checked XML containing the Gospel, the first part of the New Testament. A resulting sentence, annotated in the UD conventions, is shown in Figure 2.

Simultaneously with the replacement of the word form in Latin characters by the word form in Cyrillic characters, the linguists confronted the transcription with the printed book and they signaled the errors of the Chisinau researchers. They encountered more problems:

- A first 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 similar 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 of each word.

- Therefore, in the printed old books, these numbers are in a separated column and are written with old Cyrillic letters. These signs represent the first paragraph, the second, the third, and so on. (See Figure 4, second row). The list of letters-digits marked by an upper score has been introduced in the OCR program. The letters-digits appear only in the oldest books.

- The OCR must include a splitter that can cross out with spaces the words where they are not separated in the originally printed text, using a big lexicon for old Romanian. For example, in Figure 4, sixth row, there is a string of signs that can be separated, being a preposition, dein, =de in, En: from and,


5 Some Transcription Difficulties

The second step of the OCR processing created at Chisinau is the transliteration of old Cyrillic letters as Latin characters. The Chisinau researchers have built 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 Chisinau researchers with the second edition of the New Testament (used in the automatic morphological and syntactic annotation of the first 3,000 sentences) with Latin characters, obtained by the ABBYY FineReader 12 program. (This program is the starting point for building the OCR for Romanian Old Cyrillic and the specialists from ABBYY helped them with some indications and suggestions.) A fragment of the comparison is shown in Table 1, most of it corresponding to the image (See figure 2).

The bold characters of the table are missing words in the second edition or in the OCR version. The high quality of the first document received from Chisinau is impressive, owing mainly to the careful manual correction of OCR output. The previous examples shows that it was a very difficult task. Some difficulties of transposition of a Cyrillic letter as a group of Latin letters (with more options) are shown in the Figure 4, last row.

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 Stefan 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 pronouns co-referential with the noun of the Divinity.

Spelling correction is made for the use of a wider 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 lex-

---

**muierea** < Lat. mulier. En: the woman. This is a specific peculiarity 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, or when orthography respected any rule which we do not know.

- The book contains a lot of abbreviated words. The Chisinau researchers made a list of the individual abbreviations and they wanted the program to replace the abbreviations by 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. See Figure 4, seventh row.

- Frequently, there are no abbreviated words, but words with some overwritten letters. For example, the word *mearsără* En: they walk; the letter “r“ is overwritten; the word *Esrom* (proper noun) has an accent and the letter “m“ placed over the letter o. see Figure 4, third and fourth row. A solution applied by the OCR authors is to consider each line of text as two lines and to read the overwritten signs, too. But this peculiarity will be a permanent source for decreasing the accuracy percentage and these words must be manually checked.

- 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 white, as in Figure 3) and by increasing the resolution. The word Aminadav (proper noun) wasnt entirely recognized by the OCR program reading the black white text, but was recognized reading the color print. See Figure 4, fifth row.

The problem of letters-digit and overwritten is now solved. The output of a book treated by the OCR has some sections, like an invisible table, resembling the printed book that has a division for the notes, the number of chapters, and references of other chapters. In the Figure 3 we have manually traced the table (See figure 3)
| English Translation (KJV, Matthew 1:1-17) |
|----------------------------------------|
| The generation and birth of Jesus Christ, who is the Messiah, promised the savior to the fathers. |

| The New Testament (1648), First Printed in Romanian by Simion Ştefan, Bishop of Transylvania, obtained by OCR at IMCS Chişinău |
|--------------------------------------------------|
| The New Testament (1648), Second Edition, with Latin Letters, Alba Iulia, Romanian Orthodox Diocese Publisher, 1988 |

| Neamul și nasterea lui Iisus Hristos carele iaste Mesia făgăduit izbăvitor părinților. |
|--------------------------------------------------|
| 1. Cartea de neamul lui Iisus Hristos, fiul lui David, fiul lui Avraam. |

| Neamul și nasterea a lui is hS, carele iaste mesia făgăduit izbăvitor părinților. |
|--------------------------------------------------|
| 1. Cartea de neamul lui IS hS, fiul lui david fiul lui avraam. |

| 2. Avraam născu pre Isaac, iară Isaac născu pre Iacov, iară Iacov născu pre Iuda și pre fratei lui. |
|--------------------------------------------------|
| 2. Avraam născu pre Isaac, iară Isaac născu pre Iacov, iară Iacov născu pre Iuda și pre fratei lui. |

| 3. Iuda născu pre Fares și pre Zara din Tamar, iară Fares născu pre Esrom și Esrom născu pre Aram. |
|--------------------------------------------------|
| 3. Iuda născu pre Fares și pre Zara din Tamar, iară Fares născu pre Esrom și Esrom născu pre Aram. |

| 4. Aram născu pre Aminadav, iară Aminadav născu pre Nasson, Nasson născu pre Salomon. |
|--------------------------------------------------|
| 4. Aram născu pre Aminadav, iară Aminadav născu pre Nasson, Nasson născu pre Salomon. |

| 5. Salomon născu pre Vooz din Rahav, iară Vooz născu pre Ovid din Rutu, Ovid născu pre Iesi. |
|--------------------------------------------------|
| 5. Salomon născu pre Vooz din Rahav, iară Vooz născu pre Ovid din Rutu, Ovid născu pre Iesi. |

| 6. Iesetu născu pre David craiu, iară David craiu născu pre Solomon, den muiarea carea au fost a Uriei. |
|--------------------------------------------------|
| 6. Iesetu născu pre David craiu, iară David craiu născu pre Solomon, den muiarea carea au fost a Uriei. |

| 7. Solomon născu pre Rovoam, iară Rovoam născu pre Avia și Avia născu pre Assa. |
|--------------------------------------------------|
| 7. Solomon născu pre Rovoam, iară Rovoam născu pre Avia și Avia născu pre Assa. |

| 8. Assa născu pre Iosafat, iară Iosafat născu pre Ioaram și Ioaram născu pre Ozia. |
|--------------------------------------------------|
| 8. Assa născu pre Iosafat, iară Iosafat născu pre Ioaram și Ioaram născu pre Ozia. |

| 9. Ozia născu pre Ioatam, iară Ioatam născu pre Ahaz, iară Ahaz născu pre Ezechiiia. |
|--------------------------------------------------|
| 9. Ozia născu pre Ioatam, iară Ioatam născu pre Ahaz, iară Ahaz născu pre Ezechiiia. |

| 10. Ezechiiia născu pre Manasia, iară Manasia născu pre Amon. |
|--------------------------------------------------|
| 10. Ezechiiia născu pre Manasia, iară Manasia născu pre Amon. |

| 11. Amon născu pre Iosian, iară Iosian născu pre Ehonian și pre frații lui, în vavilon. |
|--------------------------------------------------|
| 11. Amon născu pre Iosian, iară Iosian născu pre Ehonian și pre frații lui, în vavilon. |

| 12. Iară după mutarea în Vavilon, Ehonia născu pre Salatiil, iară Salatiil născu pre Zorovavel. |
|--------------------------------------------------|
| 12. Iară după mutarea în Vavilon, Ehonia născu pre Salatiil, iară Salatiil născu pre Zorovavel. |

| 13. Zorovavel născu pre Aviu, iară Aviu născu pre Eleachim, iară Eleachim născu pre Azor. |
|--------------------------------------------------|
| 13. Zorovavel născu pre Aviu, iară Aviu născu pre Eleachim, iară Eleachim născu pre Azor. |

| 14. Azor născu pre Sadoc, iară Sadoc născu pre Achhim, Achhim născu pre Eliud. |
|--------------------------------------------------|
| 14. Azor născu pre Sadoc, iară Sadoc născu pre Achhim, Achhim născu pre Eliud. |

| 15. Eliud născu pre Eleazar, Eleazar născu pre Mattan, iară Mattan născu pre Iacov. |
|--------------------------------------------------|
| 15. Eliud născu pre Eleazar, Eleazar născu pre Mattan, iară Mattan născu pre Iacov. |

| 16. Iacov născu pre Iosif, bărbatul Mariei, dintrucat născu Iisus, carele Să chiamă Hristos. |
|--------------------------------------------------|
| 16. Iacov născu pre Iosif, bărbatul Mariei, dintrucat născu Iisus, carele Să chiamă Hristos. |

| 17. Derept acea, toate neamurile de la Avraam până la David, paturşprăzece neamare, și de la David până la mutarea în Vavilon, neamare paturşprăzece, și den mutarea den Vavilon până la Hristos, neamare 14. |
|--------------------------------------------------|
| 17. Derept acea, toate neamurile de la Avraam până la David, paturşprăzece neamare, și de la David până la mutarea în Vavilon, neamare paturşprăzece, și den mutarea den Vavilon până la Hristos, neamare 14. |
icon of processing tools all the forms existing in
the old language, in order for them to be recog-
nized and correctly analyzed when found in other
texts. For example, if we introduce in the lexicon
only capitalized proper nouns, the ones encoun-
tered without capitalization will not be correctly
annotated.

It is also worth noting that, at that time, the or-
thographic convention of words joined by a hy-
phen did not exist. For example, află-se, se află
(En: there was.) Therefore, we
must annotate two words because the string has
the meaning of two words. The inversion of word
marks, not used 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.

Another big problem is the variety of ancient
texts. At different times, in the nineteenth century,
there were transition alphabets that mixed Latin
letters with Cyrillic letters (Mărânduc et al., 2016).
16 types of transition alphabets were found, clas-
sified after the combinations of Latin and Cyril-
lic letters. There were also differences in the
form of letters, and the researchers included in
the program an option in which the user can spec-
ify the year and the locality where the text was
printed. In the sixteenth and seventeenth centuries
there were not many locations where books were
printed. They introduced in the OCR program dif-
ferent sets of letters, for each of these transitions
alphabets and printers.

6 Conclusions and Future Work

After taking the necessary steps for 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 tran-
scriptions, (both without the intention to actual-
ize, interpret or correct them), because the text is
ancient and difficult. Each of the versions com-
pared has mistaken that the other does not have,
and problems better solved than the other. The
Chisinau version is as good as the published one.
In the meantime, they have solved the problem of
the letters digits, which are missing in their ver-
sion of the table, as can be seen in Figure 3. Any
problem that linguists are telling them is solved
by the computer scientists in Chisinau. They have
reported to ASCHI the characters that are not yet
recognized and we are expecting their introduction
into international codes. They have also made a
set of old Cyrillic letters for Romanian that can be
entered on the computer keyboard, and in this way
the supervision of results of the OCR will be easi-
er.

The last 2,200 sentences of the Gospel are ob-
tained by the OCR program, and then processed in
the same way. And for the second part of the New
Testament, i.e. Acts of the Apostles, we intend
to introduce the text obtained by OCR in the pro-
grams for the automatic morphological and syn-
tactic 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 their lexicon, and by the increase
of the training corpus. The OCR will be trained
on more printed books with a higher resolution.

We will continue to wrap the gold corpus of Old
Romanian with the word form written in Cyrillic
letters. We now have 3,000 sentences, we would
need more for the training with the Cyrillic vari-
ant, to build a POS-tagger which should be able
to annotate Romanian ancient texts written in the
Cyrillic and Latin alphabets. Perhaps some Ro-
manian 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 Chisinau researchers began the study of the
problem of the optical character recognition for
old Cyrillic letters from some manuscripts. The
treebank has now three layers (conventions of an-
notation). We first annotate each text in the classi-
cal syntactic convention that our tools are trained
in, and that contains a big quantity of information.
It can be automatically transposed in the UD con-
vention, which is the international link between
more than 30 treebanks, but has less information.
Also the classical syntactic annotation can be au-
tomatically transposed in the semantic treebank
in proportion of 65%, but the syntactic relation
that have a big number of semantic correspondents
must be manually annotated. We have transposed
in the semantic format the first part (the Gospel)
of Alba Iulia New Testament, with the intention
to train a semantic parser on this format. The se-
matic annotated corpus could be aligned with the
other old New Testaments in the PROIEL project,
and they can import our semantic annotation.
The annotation of the second part of the Alba Iulia New Testament will begin in the coming year. After the full processing and supervision, the New Testament will be transformed in the UD convention and in the CONLLU format and will be affiliated to the PROEL project, which will align it with the other old New Testament versions and the linguists can study the etymology, the translation of old texts, the pragmatic connections between the sentences and other 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.

References

Lars Borin, Marcus Forsberg, and Dimitrios Kokkinakis. 2010. Diabase: Towards a diachronic blank in support of historical studies. In *Proceedings of LREC*. pages 35–42.

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of old swedish. In *LREC 2008 - Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008) Conference Proceedings*. pages 9–16.

Svetlana Cojocaru, Alexandru Colesnicov, and Ludmila Malahov. 2017. Digitization of old romanian texts printed in the cyrillic script. In *Proceedings of International Conference on Digital Access to Textual Cultural Heritage*. pages 143–148.

Alexandru Colesnicov, Ludmila Malahov, and Tudor Bumbu. 2016. Digitization of romanian printed texts of the 17th century. In *Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language*. Alexandru Ioan Cuza University Press, pages 1–11.

Mark Davies. 2010. Creating useful historical corpora: A comparison of corde, the corpus del espanol, and the corpus do português pages 137–166.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400 million word corpus of historical american english. *Corpora* (7):121–157.

Stephanie Dipper, Lukas Faulstich, Ulf Leser, and Anke Ludeling. 2010. Challenges in modelling a richly annotated diachronic corpus of german. In *Proceedings of the Conference on Language Resources and Evaluation, LREC2010*.

Wlodzimierz Gruszczynski and Maciej Ogrodniczek. 2015. The electronic corpus of the 17th and 18th century polish texts (up to 1772) aims, methods, current state, problems and prospects for development. In *Slavic Corpus Linguistics: The Historical Dimension*. pages 21–25.

Dag T. T. Haug and Marius L. Johndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data*. C. Sporleder and K. Ribarov (eds.), pages 27–34.

Cătălina Mărănduc, Ludmila Malahov, Cenel-Augusto Perez, and Alexandru Colesnicov. 2016. Rodia project of a regional and historical corpus for romanian. In *in Proceedings of MFOI*.

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

France Martineau, Constana Rodica Diaconescu, , and Paul Hirschbühler. 2007. Le corpus voies du français : de lelaboration à lannotation pages 121–142.

Tony McEnery, Richard Xiao, and Y. Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. Routledge publisher, London, New York.

C. Onelli, D. Proietti, and C. Seidenari. 2006. The diaCoris project: a diachronic corpus of written italian. In *Proceedings of LREC 2006*. pages 1212–1215. http://corpora.dslo.unibo.it/DiaCORIS.

Achim Stein. 2008. Syntactic annotation of old french text corpora. *Corpus* (7):157–172.

Irma Taavitsainen, Andreas H. Jucker, and Jukka Tuominen. 2014. *Diachronic Corpus Pragmatics*. John Benjamins publisher, Amsterdam.

Nuria Yanez-Bouza. 2011. Archer past and present (1990-2010). *ICAME Journal* (35):205–236.