



**A DIACHRONIC CORPUS FOR  
ROMANIAN (RODIA)**

8-10 MARCH 2017, DGFS-SAARBRÜCKEN

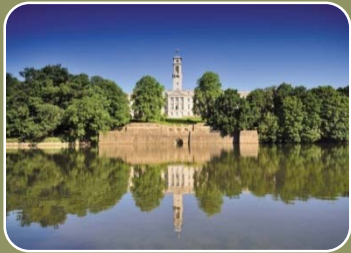
*Cătălina Mărănduc<sup>1</sup>, Ceneș-Augusto  
Perez<sup>1</sup>, Ludmila Malahov<sup>2</sup>, Alexandru  
Colesnicov<sup>2</sup>*

<sup>1</sup>Faculty of Computer Science, Al. I. Cuza University, Iași, Romania

<sup>2</sup>Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova



Beginning with 2005, a lot of corpora for old languages are built: for Spanish, Portuguese, Italian, but also for German, Polish, English, Dutch, Swedish, Finnish, Estonian, Japanese, and corpora for dead languages.



Conferences for the diacronic corpora processing:

- [Http://palc.Uni.Lodz.Pl/](http://palc.Uni.Lodz.Pl/)
- <https://www.Nottingham.Ac.Uk/conference/fac-arts/clas/dcgic/home.aspx/>



<http://dcs1.orinst.ox.ac.uk/>

<http://bvmcresearch.cervantesvirtual.com/diasearchtool/>

<http://www.corpusdelespanol.org/x.asp/>

<http://www.corpusdoportugues.org/x.asp/>

<http://corpus.byu.edu/historical-syntax.asp/>

# RELATED WORK

- ❖ The diachronic corpus of Italian is described in "The DiaCORIS project: a diachronic corpus of written Italian" by C. Onelli, D. Proietti, C. Seidenari (LREC 2006).
- ❖ There are 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: <http://dcs1.orinst.ox.ac.uk/>
- ❖ The creation of a German diachronic corpus is described in "Challenges in Modelling a Richly Annotated Diachronic Corpus of German" by S. Dipper, L. Faulstich, U. Leser, A. Ludeling (LREC 2010).
- ❖ A comparison between the historical Spanish and Portuguese corpora is made in "Creating Useful Historical Corpora: A Comparison of CORDE, the Corpus del Español, and the Corpus do Português" by M. Davies, in *Diacronía de las Lenguas Iberorromances: Nuevas Perspectivas desde la Lingüística de Corpus*, ed. Andrés Enrique-Arias, 2010.
- ❖ In "Something Old, Something New: A Computational Morphological Description of Old Swedish" made by L. Borin, and M. Forsberg (LREC 2008).

# UAIC-RO-DEP-TB

- ❖ Dependency Treebank for Romanian, built at the Al. I. Cuza University of Iasi.
- ❖ Balanced corpus: Standard language (legal style frame-net, wikipedia, quotations, 1984 Novel)
- ❖ Non-standard:
  - Social-media: 2,503 sentences, 39,290 words
  - 17-th century: 3,482 sentences, 58,103 words
  - Popular and regional: 230 sentences, (and 2,000 sentences, in work)
  - The Novel "The Kings of Old Court" (1910) -1,650 sentences, 53,000 words
- ❖ Total (standard & non-standard): 15,400 sentences, 301,139 words.



# MORPHOLOGICAL LEVEL

- ❖ Starting with the Multext East set of labels, a big amount of information.
- ❖ We gave up the clitics annotation - in the old language an infinite number of changes caused by the hyphen may occur.
- ❖ We refined the annotation of verbs - with the intention of building a dictionary of verbal patterns:
  - The participle is always annotated as verb, and not as adjective
  - The negative form of impersonals and imperatives
- ❖ We have 500 morphological labels

## EXAMPLES:

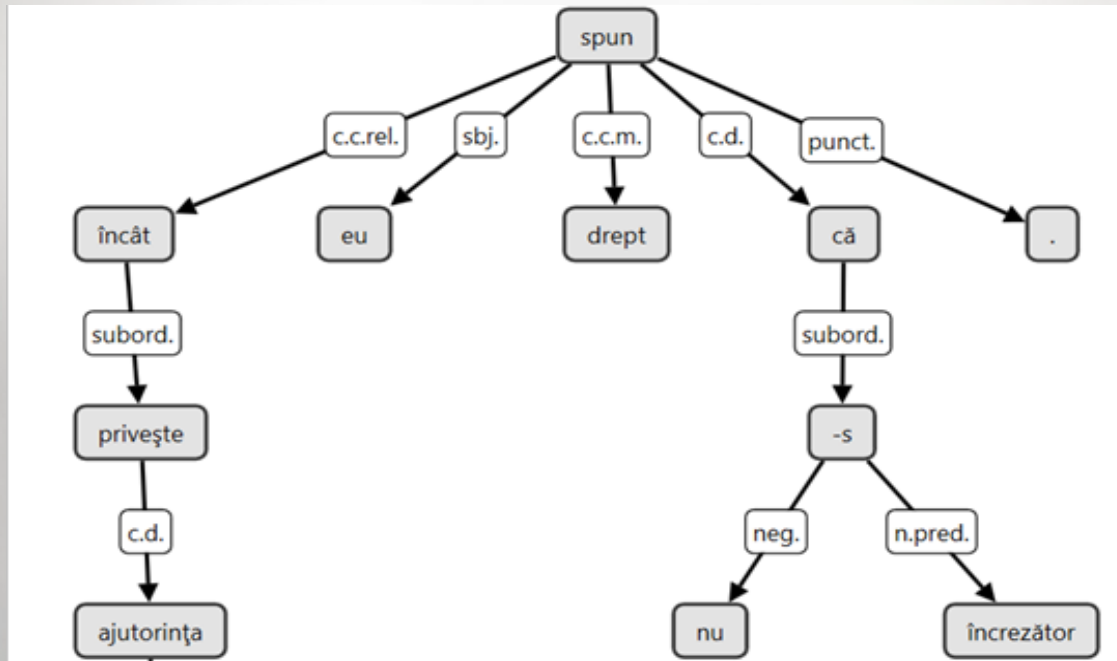
- ❖ Words as "neterminat", "nesticat", "**neștiind**" (En: unfinished, unbroken, not knowing)
- ❖ have lemma "termina", "strica", "**ști**",
- ❖ the verbs "\*a netermina", "\*a nesticat", "\*a **nești**" (En: \*to unfinish, \*to unbreak, \*to unknow) do not exist.
- ❖ The postag of these forms will be annotated as: "**Vmp--sm-z**", "**Vmp--pf-z**", "**Vmg----z**". (participle negative singular masculine, participle negative plural feminine, gerund negative).
- ❖ The opposite label to **Vmp--pf-z** has on the eighth position the p that means "positive";
- ❖ "nesticat" = **Vmp--pf-p**.

# SYNTACTIC CONVENTIONS OF ANNOTATION

- ❖ Classical syntactic labels, such as: c.c.l., c.c.t. – 14 types of circumstantial modifiers, c.ag. – agent complement, refl. – mark of reflexive diathesis.
- ❖ Coordination as oblique subordination starting with the first element.
- ❖ The relative words-tools are heads for the function which they introduce.
- ❖ 44 syntactic relations.
- ❖ Our labels contain more information than UD ones (universal dependencies).
- ❖ It is important to keep unchanged the annotation conventions in order to increase the consistent training corpus for the parser.

# EXAMPLE

❖ As regards the help as I say I'm not overconfident.





# GOALS (I)

- ❖ To use the morphological annotation for building big lexicons (with a big amount of lexical variants) for the Old-Ro-POS-tagger and for the Old-Ro-Cyrillic-OCR program.
  - We extract the variants from the Thesaurus Dictionary.
  - We carefully check the automatic morphological anntotation.
  - We extract the checked information from the Treebank to be introduced in the lexicon.

## GOALS (II)

- ❖ To use the classical syntactic annotation (being rich in information) as pivot for the transformation in other conventions of annotation.
  - Universal Dependency
  - PROIEL (aims to align the oldest Latin, Greek, Slavonic and Armenian New Testaments)
  - Semantic annotation
- ❖ To build tools for the Old Romanian processing, based on the UAIC tools (or Chisineu tools) for Contemporary Romanian.

## EXAMPLE 1: COLLECTING LEXICAL VARIANTS

	form	lemma	postag
DOMESTICÍE ( <b>lemma</b> ) s. f. ( <b>postag</b> ) - Și: domesnicíe, dumesnicíe s. f.	domesnicie dumesnicie	domesticie_Ncfsrn domesticie_Ncfsrn	
DOMESTICÍRE s. f. - Și: (învechit) domesnicíre (polizu, LM), dumesnicíre s. f.	domesnicire dumesnicire	domesticire_Ncfsrn domesticire_Ncfsrn	
DOMESTICIT, -Ă adj. - Și: (învechit și regional) dumesnicít, -ă, (învechit) domesnicít, -ă, dumesnicít, -ă, dumestnicít, -ă, (învechit, rar) domesticát, -ă (LM) adj.	dumesnicit dumesnicită dumesnicit dumesnicită dumestnicit dumestnicită dumesticat dumesticată	domestici_Vmp--sm-p domestici_Vmp--sf-p domestici_Vmp--sm-p domestici_Vmp--sf-p domestici_Vmp--sm-p domestici_Vmp--sf-p domestici_Vmp--sm-p domestici_Vmp--sf-p	

## EXAMPLE 2: EXTRACTING DATA FROM THE TREEBANK

Ce	LEMMA=	ce	MSD=	Pw3--r
ce	LEMMA=	ce	MSD=	Pw3--r
ceaste	LEMMA=	acest	MSD=	Dd3fpr
cel	LEMMA=	cel	MSD=	Tdmsr
cel	LEMMA=	cel	MSD=	Pd3msr
ceriu	LEMMA=	cer	MSD=	Ncmsrn
ceriurelor	LEMMA=	cer	MSD=	Ncfpoy
ceriurile	LEMMA=	cer	MSD=	Ncfpry
cetate	LEMMA=	cetate	MSD=	Ncfsrn
cetatea	LEMMA=	cetate	MSD=	Ncfsry
chema	LEMMA=	chema	MSD=	Vmn
chema	LEMMA=	chema	MSD=	Vmii3s



# HISTORY

- ❖ The existence of the books printed in the old Cyrillic alphabet (in Romania before the middle of the nineteenth century, în Moldova before 1990), is a common problem for Romania and the Republic of Moldova, two countries where the Romanian language is spoken.
- ❖ Therefore, the researchers in NLP from the two countries decide to solve them together.
- ❖ Both in Romania and in Republic of Moldova, the old texts, the first printed books are written in an old Cyrillic alphabet that has 47 letters and is not recognized as the ASCII encoding. Some of the corresponding Unicode points were introduced only since 2009, and we found only several fonts covering them: ↑ 𐌀, 𐌁.
- ❖ In the nineteenth century, a lot of transition alphabets occurred.

# CYRILLIC AND TRANSITION ALPHABETS

Фѣнлѡр Ѣншпектѡрѣ, Ѣдминистраѡрѣ,  
 Нотарѣшнѣ, Парѡхнѣ шн Прѣшнѣ! Дар вѡлѡ  
 шн паче дела милостнѣва Дмнезѣѡ гѡрѡ  
 дела Нѡнѣ Ѣрхїерѣска благаблѡженїѣ .

Bobb·1808

кѡ аст кїп сѡ спарїѣ пе ачеї каре ар  
 ѡдрѣснї сѡ тѡ жѡдече. Дар ої idee рѡ-  
 нїде тѡ фѡкѡ сѡ'мї скїтѡ хотѡрїреѡ:  
 тѡ гѡндїїѣ кѡ о асемене колекцїе пѡ

Alexandrescu·  
1838

A text with transition alphabet, the letters: *a*, *d*, *n*, *i*, *e*, *m* are Latin

Anatomia se пумече зоологїкѡ саѡ ком-  
 паратѡ, кѡнд ѡнѡ'шн студїѣ генерал, се ѡтврѣ-  
 цїшѡсѡ tot шїрѡ добитѡчелѡр, черчетѡндѡсе шї

Kretzulescu·  
1843

Anoter transition alphabet, the Latin letters *t*, *s*, *z* were added to the above ones

## THE OCR PROGRAM FOR THE OLD CYRILLIC BOOKS

- ❖ It is the first time we have Romanian texts with editable old cyrillic letters!
- ❖ What OCR is needed for?
- ❖ We have a lot of editions of texts which the specialists manually wrote with Latin letters after looking at images. The editions include: comments, texts and final indexes.
- ❖ But the text is written in interpretative spelling (according to controvertible theories of the sixteenth and seventeenth century's pronunciation and transcription agreed by the editor), and indexes contain the lemma, actual form of word, with pages numbers where the old form can be found.
- ❖ If we include words of these texts or the indexes in the POS-tagger lexicon, it will not found them in the old text analysed.
- ❖ The non specialists probably will not read these books even if they are updated, easier to undersand.
- ❖ The specialists are interested in true real form of old texts and cannot access it in the commented new editions.

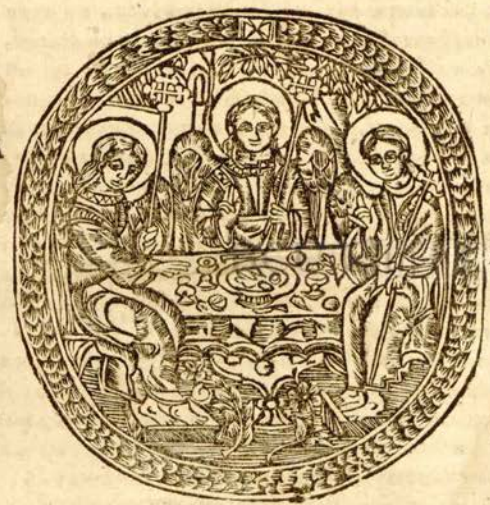


## **DIGITISATION MEANS NOT ONLY PRINTING, BUT ALSO READING OLD BOOKS**

- ❖ In Romanian, the old books of the big libraries were scanned by Dacoromanica, the Soros foundation, in the Republic of Moldova by Moldavica. More than 100 old books were downloaded to train the OCR for old Romanian.
- ❖ <https://www.google.ro/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=dacoromanica.ro>
- ❖ <http://www.moldavica.bnrm.md/index.html>
- ❖ We chose the first New Testament printed in Romanian (Alba Iulia 1648) with the intention of aligning the project PROIEL. The book is well conserved, scanned by the University of Cluj-Napoca.



Прѣ лѣмнѣтѣл Сѣмнѣ, ал СФѣтѣи шѣ не дѣспрѣцѣтѣи  
Тронцѣ, Храмѣлѣ, Митрополѣи Бѣлградѣлѣи,  
днѣ Ардѣлѣ.



ГЛАВА ІТѢИ БДНОІЩІИИ ИЖИКОТВОРАЩІИ ИИ  
РАЗДѢЛИМѢИ ТРОИЦѢ, ШЦѢИ ИИШІИ ІТѢМѢ ДѢШ,  
ВЪКІГДА ИИѢ И ПУИИИИ И ВЪЗѢКІИ  
ВЪКОМѢ, АМІИИИ.

ПРЕДОСЛОВІЕ КЪ ТРЪЖ ЧЕТИТОРОИ.




Четиторилорѣ, Архидѣлѣи СФѣтѣи Кѣрѣ, мнѣж  
пѣи шѣ гнѣтѣтѣ, дѣлѣ Тѣтѣлѣ, Дѣмнѣлѣ,  
нѣтрѣ ІС ХС.

Дѣспрѣцѣи, дѣспрѣтѣтѣ лѣсрѣрѣи Кѣрѣлѣ гнѣтѣ иерѣи прѣрѣдѣ.

**И**ИТ ТѣтѣлѣиТѣ лѣлѣ Архидѣлѣи ал ИЖВОДИ, ЕРМОНѢ  
Сѣлнѣтѣрѣ, днѣ порѣнкѣ шѣ кѣлѣшѣлѣлѣ мѣрѣи  
лѣлѣ. шѣ сѣлѣ лѣлѣ ІЩІИИИТѣ Кѣтѣ лѣлѣ Пѣтѣтѣ. шѣ  
Кѣрѣтѣ Архидѣлѣ Тѣтѣлѣтѣ лѣлѣ мѣлѣтѣ. шѣрѣ нѣ ІКОТѢИ ДѢШѢРѢ А  
лѣлѣ мнѣтѣ гнѣтѣлѣ мѣлѣтѣ лѣлѣ шѣ, шѣ грѣшѣлѣ Архидѣлѣтѣрѣ  
лѣлѣ Пѣтѣтѣ нѣ Архидѣлѣлѣ лѣлѣ мѣтѣи шѣ Кѣрѣи грѣици. Пѣи  
тѣрѣ тѣлѣ нѣнѣ лѣлѣ Архидѣлѣ днѣтѣлѣ алѣ Пѣлѣлѣшѣи шѣ ІЩІИ  
дѣ нѣлѣ Фѣтѣтѣ бнѣи лѣлѣ ИПРѣВѢИТѢ шѣ лѣлѣ Архидѣлѣтѣ. шѣ  
лѣлѣ ТОКМѢИТѢ днѣ Архидѣлѣтѣ лѣлѣ Пѣтѣтѣ.

Чѣ ИЖМАИ Архидѣлѣтѣ шѣици, Кѣ нѣнѣ нѣлѣмѣ ІКОТѢИТѢ НѣМАИ  
прѣ ІЩІИ ИЖВОДИ ТѣТОЛѢ Кѣтѣ Архидѣлѣтѣ Архидѣлѣ, грѣици, шѣ  
сѣрѣици, шѣ лѣтѣици, Кѣрѣлѣ Архидѣлѣтѣ Фѣтѣтѣ ИЖВОДИТѢ Дѣ  
Кѣрѣтѣлѣлѣи мѣлѣи шѣ Архидѣлѣтѣтѣи Архидѣлѣтѣ грѣици, лѣлѣ  
лѣтѣтѣ шѣ лѣлѣмѣ ІКОТѢИТѢ, тѣ мѣлѣи ВЪРѢТОСѢ ИТѢМѢ ЦИИИТѢ  
дѣ ИЖВОДИЛѢ грѣици. шѣ лѣлѣ ІКОТѢИТѢ шѣ прѣ ИЖВОДИЛѢ  
лѣлѣ ВРОНИИИИИИ, Кѣрѣлѣ Архидѣлѣтѣ днѣ Тѣлѣ, днѣ лѣлѣ мѣлѣ  
грѣици, лѣтѣици, шѣ лѣлѣ ІКОТѢИТѢ шѣ ИЖВОДИЛѢ  
сѣлѣици Кѣрѣи ИЖВОДИТѢ ІКОТѢИТѢ, днѣ грѣици, шѣ  
шѣ Тѣтѣлѣтѣ Архидѣлѣ Мѣлѣлѣшѣи. шѣ ІКОТѢИТѢ Архидѣлѣ  
Тѣлѣтѣ Вѣрѣ Кѣрѣтѣ Архидѣлѣтѣ мѣлѣ Архидѣлѣ дѣ Кѣрѣтѣ грѣи  
ци дѣ прѣ Архидѣлѣ ІКОТѢИТѢ, Архидѣлѣ дѣ грѣици нѣ  
ИТѢМѢ Дѣ Пѣтѣтѣ. шѣици Кѣ Архидѣлѣтѣ ІЩІИТѢ Архидѣлѣ  
НАТѢ БѢЛѢИТѢИ, шѣ

## BUILDING THE LEXICON FOR THE OCR PROGRAM

- ❖ The set of letters used was Times extro, except the three letters that are not recognised by the universal programs.
- ❖ The sign  was replaced with **an arrow "↑"**;
- ❖  was maintained in the form that has been used by **the researchers of Chisinau, "щ", although this letter is spelled şc, and not şt, as it should be.**
- ❖ For the letter  a quite similar character was found, "8".
- ❖ With this set of letters, we replaced the word form in the xml automatically morpho-syntactic annotated, then checked, in order to obtain a lexicon for the OCR program.



# EXAMPLE:

```
<sentence id="32" parser="Radu's parser" user="ugla" date="2016-4-16" citation-part="MATT.2.14.content">
  <word id="1" form="ѡѡ" lemma="u" postag="Tf-so" head="2" chunk="" deprel="det"/>
  <word id="2" form="ѠѡснѠ" lemma="Iosif" postag="Npmsm" head="3" chunk="" deprel="iobj"/>
  <word id="3" form="ѡѡѡѡѡѡ" lemma="porunci" postag="Vmis3s" head="0" chunk="" />
  <word id="4" form="Ѡѡѡѡѡ" lemma="inger" postag="Ncmsry" head="3" chunk="" deprel="nsubj"/>
  <word id="5" form="ѡѡѡ" lemma="si" postag="Ccssp" head="3" chunk="" deprel="cc"/>
  <word id="6" form="ѡѡѡѡ" lemma="fugi" postag="Vmis3s" head="3" chunk="" deprel="conj"/>
  <word id="7" form="ѡѡѡ" lemma="cu" postag="Spsa" head="8" chunk="" deprel="case"/>
  <word id="8" form="Ѡѡѡ" lemma="Iisus" postag="Npmsm" head="6" chunk="" deprel="nmod"/>
  <word id="9" form="ѡѡѡ" lemma="si" postag="Ccssp" head="8" chunk="" deprel="cc"/>
  <word id="10" form="ѡѡѡ" lemma="cu" postag="Spsa" head="11" chunk="" deprel="case"/>
  <word id="11" form="ѡѡѡѡѡ" lemma="mamă" postag="Ncfsry" head="8" chunk="" deprel="conj"/>
  <word id="12" form="ѡѡѡ" lemma="el" postag="Pp3mso" head="11" chunk="" deprel="nmod"/>
  <word id="15" form="." lemma="." postag="PERIOD" head="3" chunk="" deprel="punct"/>
</sentence>
```

❖ The angel commanded to Joseph and he flee with Jesus and his mother. (Alba Iulia New Testament)

## EXTRACTING THE INFORMATION FROM THE XML

❖ The CONLLU format of another sentence from the Alba Iulia New

1	→	<u>Нѣмѣл</u> ·neam	→		→	<u>Nemsry</u>	→	0	→	<u>ROOT</u>	→		→					
2	→	<u>ши</u>	→	<u>și</u>	→		→		→	<u>Ccssp</u>	→	1	→	<u>coord.</u>	→			
3	→	<u>нашерѣа</u>	→	<u>naștere</u>	→		→		→	<u>Ncfsry</u>	→	2	→	<u>coord.</u>	→			
4	→	<u>алѣ</u>	→	<u>-ul</u>	→		→		→	<u>Tfmsr</u>	→	5	→	<u>det.</u>	→			
5	→	<u>Іс</u>	→	<u>Iisus</u>	→		→		→	<u>Npmsrn</u>	→	3	→	<u>a.subst.</u>	→			
6	→	<u>хс</u>	→	<u>Hristos</u>	→		→		→	<u>Npmsrn</u>	→	5	→	<u>a.subst.</u>	→			
7	→	<u>кареле</u> ·care	→		→		→		→	<u>Pw3msry</u>	→		→	5	→	<u>a.vb.</u>	→	
8	→	<u>аѣте</u>	→	<u>fi</u>	→		→		→	<u>Vmip3s</u>	→	7	→	<u>subord.</u>	→			
9	→	<u>месіа</u>	→	<u>Mesia</u>	→		→		→	<u>Npmsry</u>	→	8	→	<u>n.pred.</u>	→			
10	→	<u>фѣгѣдѣит</u>	→	<u>făgădui</u>	→		→		→	<u>Vmp--sm</u>	→		→	9	→	<u>a.vb.</u>	→	
11	→	<u>избѣвитор</u>	→	<u>izbăvitor</u>	→		→		→	<u>Afpmsrn</u>	→		→	9	→	<u>a.adj.</u>	→	
12	→	<u>пѣринцилор</u>	→	<u>părinte</u>	→		→		→	<u>Ncmroy</u>	→		→	9	→	<u>c.i.</u>	→	

Nation and Birth of Jesus Christ, who is Messiah promised, deliverer of parents.



## LEXICON FOR THE OCR (EXCERPT)

<b>маре</b>	LEMMA=	mare	MSD=	Afpmsrn
<b>маре</b>	LEMMA=	mare	MSD=	Ncfsrn
<b>марелѣи</b>	LEMMA=	mare	MSD=	Afpmsoy
<b>маріи</b>	LEMMA=	mare	MSD=	Ncmpry
<b>марѣ</b>	LEMMA=	mare	MSD=	Ncfsry
<b>маги</b>	LEMMA=	mag	MSD=	Ncmprn
<b>маг</b>	LEMMA=	mag	MSD=	Ncmsrn
<b>маи</b>	LEMMA=	mai	MSD=	Rg
<b>маи маріи</b>	LEMMA=	mai~mari	MSD=	Ncmpry
<b>маинте</b>	LEMMA=	mai~înainte	MSD=	Rg

# TRAINING AND PROCESSING

- ❖ The OCR program has been trained on texts printed after 1945 in The Republic of Moldova with the modern Cyrillic alphabet.
- ❖ 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 with old Cyrillic characters.
- ❖ The second step of the OCR processing is the transliteration of old Cyrillic letters as Latin characters.
- ❖ A POS-tagger that can annotate with morphological information both Romanian texts written with Latin letters and written with modern Cyrillic characters is used; the program was adopted for old Cyrillic (introducing new characters) and for the operation of transposition Cyrillic to Latin (introducing new rules).

# EVALUATION OF RESULTS

- ❖ The text with Latin letters obtained by the OCR program of Chisinau is compared with the second edition of the Alba Iulia New Testament, edition with Latin letters made by priests, without the intention to "actualise".
- ❖ Below, we have shown an excerpt of the comparison between the second edition and the OCR of the first edition; the red mark the mistakes or the missed words in the OCR version, or in the second edition.

The New Testament at Alba Iulia (1648)  
Second Edition, with Latin Letters, Alba  
Iulia, Publisher Romanian Orthodox  
Diocese of Alba Iulia, 1988.

**Neamul și nașterea lui Iisus Hristos carele iaste Mesia  
făgăduit izbăvitor părinților.**

1. Cartea de neamul lui Iisus Hristos, fiul lui David,  
fiul lui Avraam.
2. Avraam născu pre Isaac, iară Isaac născu pre Iacov,  
iară Iacov născu pre Iuda și pre frații lui.
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 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 a Uriei.
7. Solomon născu pre Rovoam, iară Rovoam născu pre  
Avia și Avia născu pre Assa.
8. Assa născu pre Ioasafat, iară Ioasafat născu pre  
Ioaram și Ioaram născu pre Oziia.
9. Oziia născu pre Ioatam, iară Ioatam născu pre Ahaz,  
iară Ahaz născu pre Ezechiiia.
10. Ezechiiia născu pre Manasia, iară Manasiia născu pre  
Amon.
11. Amon născu pre Iosian, iară Iosian născu pre  
Ehonian și pre **frații** lui, în vavlon.

The New Testament at Alba Iulia (1648)  
First Printed in Romanian by Simion **Ștefan**,  
Metropolitan Bishop of Transylvania.  
Ocreized at IMATH **Chișinău** 02.08.2016.

**N**Heamul și nașterea 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 frații lui .
3. Iuda născu pre fares, și pre zara din Thamar. iară fares  
născu pre Esrom, și 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,  
și avia născu pre assa.
8. Assa născu pre Iosafat. și Iosafat născu pre Ioaram. și  
Ioaram născu pre ozia.
9. Ozia născu pre Ioaθam. iară Ioaθam 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 și  
pre frații lui, în vavlon.



# OBSERVATIONS

❖ The **letter ȳ does not turn into the Latin j, but** into ge, gi. The cyrillic letter "s" is traditionally spelled dz, the letter ʁ is spelled th, and the group "ou" is always spelled "u".

❖ Examples:

❖ :↑**цѣрѣл** = îngerul, not înjjerul "the angel";

❖ **дѣмнѣсеѣ**=dumnedzeu, and not dumneseu "God",

❖ **оунде**=unde, not ounde "where".

## STUDYING AND PRESERVING PECULIARITIES OF OLD TEXT (I)

- ❖ 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.
- ❖ If we introduce in the lexicon only capitalized proper names, the ones encountered without capitalization will not be correctly annotated.
- ❖ The words joined by a hyphen did not exist. For example, ***aflăsă*** means ***află-se***, ***se află*** "there was".
- ❖ The inversion of word marks ***află-se***, not accepted in the contemporary language, is frequently used in ancient texts.

## STUDYING AND PRESERVING PECULIARITIES OF OLD TEXT (II)

- ❖ The numbers of chapters and paragraphs are in a separate column, written with numbers-letters:
  - ❖ א̣ = 1; ב̣ = 2; ג̣ = 3, and so on.
- ❖ There are words without white space between: אֶת־מִשְׁרָפֶת  
= אֶת composed preposition *de în* "from" + מִשְׁרָפֶת noun, *muiarea* "the woman". Strings as: *Cee întruia asănaște = Ce e întru ia a să naște* "what's in her to be born" seem to provide guidance on pronunciation and phrasing.
- ❖ The book contains a lot of abbreviated words. We must decode them and dress a list of abbreviations.
- ❖ There are words with some overwritten letters. For example, the word עֶרְוֹ, Esrom, proper noun, has the letter m written over the letter o.



## CONCLUSIONS AND FUTURE WORK

- ❖ It is a good idea to compare two transcriptions, (both without the intention to actualize, interpret or correct the old text); each of them can have some errors.
- ❖ Each operation will be supervised and, the programs (OCR and POS-tagger) will be ameliorated by the introduction of the correct information in the lexicon and by the training on an increased gold corpus.
- ❖ 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.

**THANK**

**YOU!**

