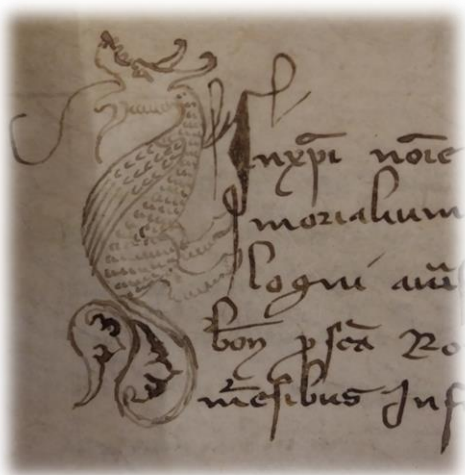


Challenges and issues of using *Transkribus* in large late medieval manuscript collections: The Memoriali Project (*MemoBo*)



Edward Loss



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PROGETTO DI RICERCA
MEMOBO

1) The *Memoriali*

Definition

The collection in the State Archive of Bologna

2) The *MemoBo* project

Aims and objectives

X-Dams, *Regesta* and the state of the art

Website layout

3) The *Memoriali* and *Transkribus*

Data input format: Diplomatic transcription + *tags* vs direct transcription

First model: *MemoBo*, *Me. 69*, *Model 1*

Second model: *MemoBo*, *Me. 69*, *Model 2*

Third Model: *MemoBo*, *Me. 69*, *Model 3*

4) Next steps and future plans

Definition:

- A series of notarial records – contracts, wills, dowry stipulations, emancipations etc – registered by a specific office, the «ufficio dei Memoriali» and created in order to avoid fraud and falsification.



- Product of a statutory (*statuta*) disposition: **April of 1265**

*Rubrica XLIII- Qualiter contractus et ultime voluntates per notarios in **memorialibus** reducantur et qualiter ipsi notarii elligantur et qualiter ipsa memorialia fiant. Statuimus et **ordinamus ut falsitatibus que circa instrumenta fiebant omnimode obvietur** quod omnes contrahentes deinceps in civitate bononie et burgis, (...) *Statuti di Bologna dall'anno 1245 all'anno 1267*, III, a cura di Luigi Frati, Bologna, 1877, pp. 625-631.*

- The series only contains documents in the value of 20 bolognese lire or more.

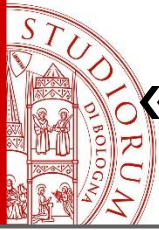


The *Memoriali* series in the State Archive of Bologna (ASBo)

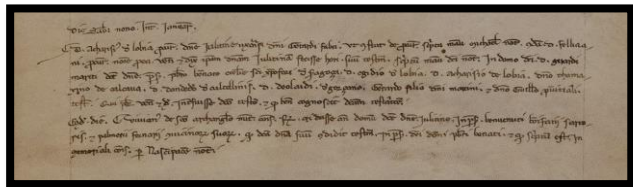
Composed of 322 volumes produced between 1265 and 1452,
organized by semester.



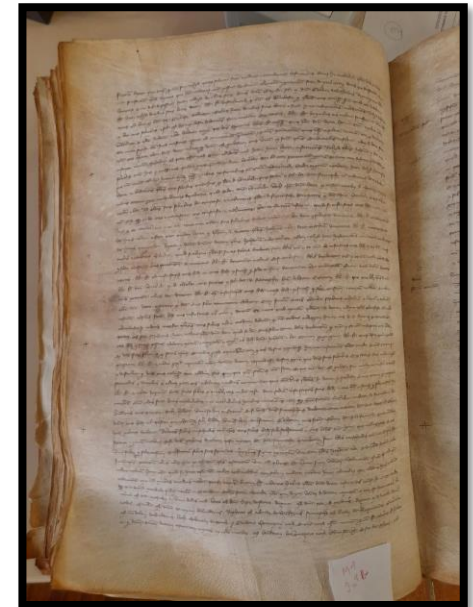
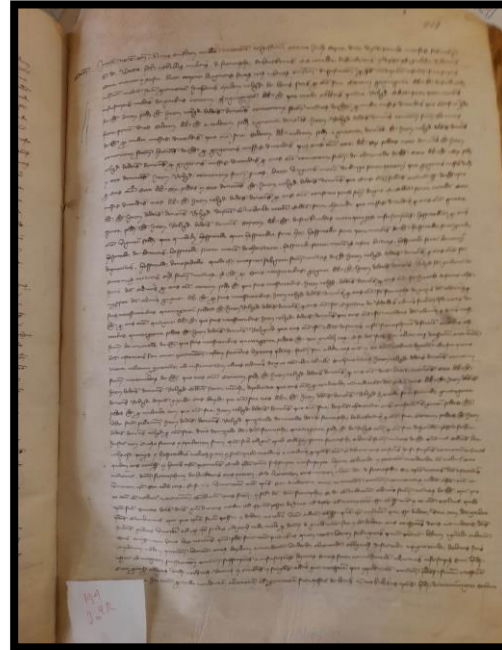
More than a million
acts uninterruptly
covering 200 years of
bolognese economic and
social activities.



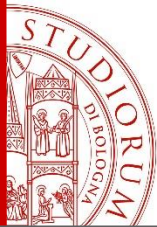
«...documents that were continuously evolving»



Will of Jubetina, wife of Gerado, blacksmith (09/01/1266)



Will of Dota, daughter of Francesco dei Clarissimi – 20/02/1338



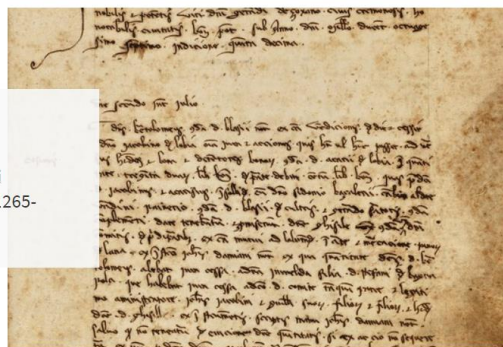
MemoBo

[HOME](#)[PROGETTO](#)[PERSONE](#)[PARTNER](#)[DATABASE](#)[AGENDA](#)[PILLOLE](#)[BIBLIOGRAFIA](#)[FOTOGALLERY](#)[LINK](#)


[Home](#) / [Progetto](#)

Progetto

MemoBo. Un mare magnum di possibilità: i Memoriali bolognesi e la loro schedatura (1265-1452)

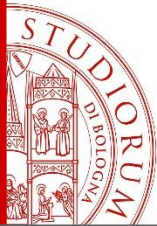


I Memoriali
bolognesi

 I libri memorialium del Comune di Bologna

L'Ufficio dei Memoriali fu istituito dal Comune di Bologna nel 1265 per registrare in

<https://site.unibo.it/memobo/it/progetto>

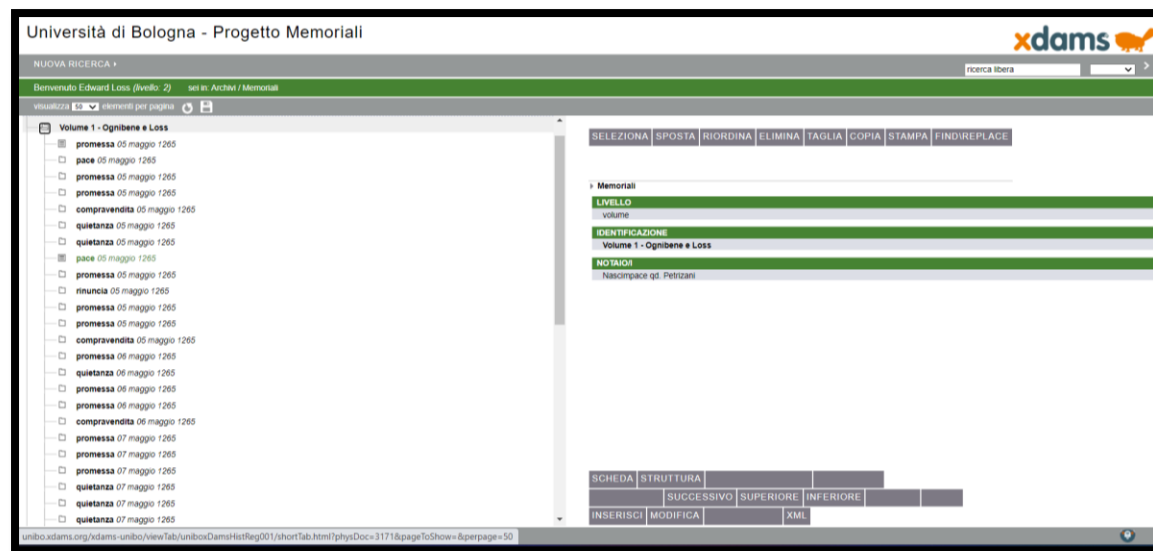


MemoBo

1) The creation of an **open-access search tool** with different access keys to the documentary series.

2) The organization of **serial data** that will provide scholars with a solid basis for more detailed research on economic, social, and legal history of a main medieval city.

- Together with *Regesta*, we created a data recording and indexing software in 2020 (on X-Dams)





MemoBo - XDams



INSERISCI

CHIUDI X

è stata selezionata la scheda: Volume 1 - Ognibene e Loss

SALVA E CHIUDI
SALVA E
CONTINUA
CHIUDI

informazioni di relazione

- ☐ figlio
☐ fratello precedente
☒ fratello successivo

+ inserimento multiplo di schede

- IDENTIFICAZIONE DELL'UNITÀ

visibilità della scheda

livello di descrizione

categoria

- DATA DI REDAZIONE

da / /

a / /

forma visualizzata

forma normalizzata

indizione

nota alla data

APPLICA

SENZA DATA

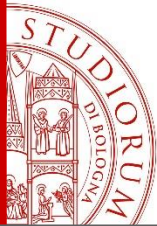
+ DATA TOPICA: LUOGHI DI CITTÀ

+ DATA TOPICA: GUARDIA CIVITATIS

+ DATA TOPICA: LUOGHI DEL CONTADO

+ DATA TOPICA: LUOGHI EXTRA GIURSDIZIONE

- NOTAIO REGISTRATORE



MemoBo - XDams



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PROGETTO DI RICERCA

MEMOBO

MODIFICA

CHIUDI X

IDENTIFICAZIONE E DESCRIZIONE | INDICI NOMI | EDIZIONI, NOTE E COMPILAZIONE

- PRECEDENTE
- SUCCESSIVO
- SUPERIORE
- SALVA
- SALVA E CHIUDI
- CHIUDI

Area indici nomi

-

PERSONE

-

AGGIUNGI

genere

...

nome

cognome

età

/c/controlaccess[@altrender='persone']/list/item[1]/persname[@altrender='cognome']/text()

provenienza o residenza

qualifica

ente di appartenenza/riferimento

APRI

condizione

mestiere

funzione

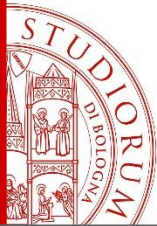
...

APRI note

INSERISCI

RIMUOVI





MemoBo - XDams

Some of the information catalogued and indexed:

- 1) Type of the act
- 2) People involved and their roles
 - 2.a) name, patronyc and surname
 - 2.b) geographical provenance
 - 2.c) profession
- Etc
- 3) The object
 - 3.1) transferable goods, rights and etc
 - 3.2) their value and type of currency

- Main challenge:
the size of the collection!

MemoBo and A. I.

The search for A. I.

- Study of the options available in the market:

1) Supply the lack of manpower

2) Possible solution for the challenge of time



RESCRIBE

eScriptorium



MemoBo and *Transkribus*

Advantages:

Easier to install

Easier to use than most softwares

Sustainable structure (more possibilities of lasting)

Disadvantages:

Private for profit endeavour: no Italian university can establish a direct partnership according to current legislation, which complicates funding.

Costs considering the size of the collection



MemoBo and Transkribus

First phase:

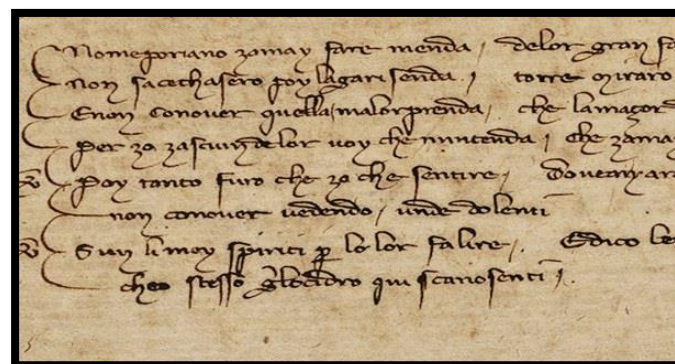
Selection of a *memoriale* to start training the software and creating specific models.

Criteria: the clarity of script, the clarity of support (*parchment*) and the existence of a (almost) complete man-made perfected transcription.

Selected:

Memoriale of Enrichetto delle Querce (1287)

- Clear thirteenth-century notarial minuscule
- Precise grammar and syntax – notary of high intellectual and cultural relevance.
- Document of high historical relevance: Dante's poems appear in his *Memoriale*
- Partially transcribed by Armando Antonelli (State Archive of Bologna)



«The Garisenda Sonet»



MemoBo and Transkribus

Models of data input for the creation of «ground truth»

1) Diplomatic transcription + «tags»

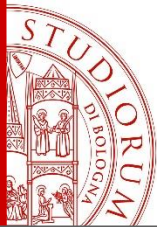
- Transcription of graphic elements as we see them on the parchment, copying medieval characters and specific abbreviation trace from MUFI (Medieval Unicode Font Initiative)
- Expansion of abbreviations through the «tag» system.

The screenshot displays the MUFI website interface. At the top, the MUFI logo and the text 'Medieval Unicode Font Initiative' are visible. Below the header, there is a navigation menu with links: Home, Browse by character, Browse by code chart, Browse by range, Browse by updates, and Full code chart. The main content area shows a grid of medieval characters and their corresponding Unicode code points. The grid is organized into rows and columns, with each cell containing a code point (e.g., 0020, 0021) and a character (e.g., space, exclamation mark). The characters are arranged in a way that shows the progression of the medieval alphabet and its various forms.

<https://mufi.info/m.php?p=muficodechart#&ui-state=dialog>

1) Diplomatic transcription + «tags»

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA





MemoBo models of data input: Diplomatic transcription + «tags»

Advantages:

- Smaller number of pages necessary to create a model
- More precise reading models (in theory)
- Especially useful for the production of critical editions

Disadvantages:

- **Time-consuming process:** 2-3 hours of preparation for each page.
- **Very arbitrary process:** the selection of characters input depends on the observer's individual understanding of the script. Es. cōē , coe or coē for ?
- Not all characters available on MUFI are supported by transkribus: «»

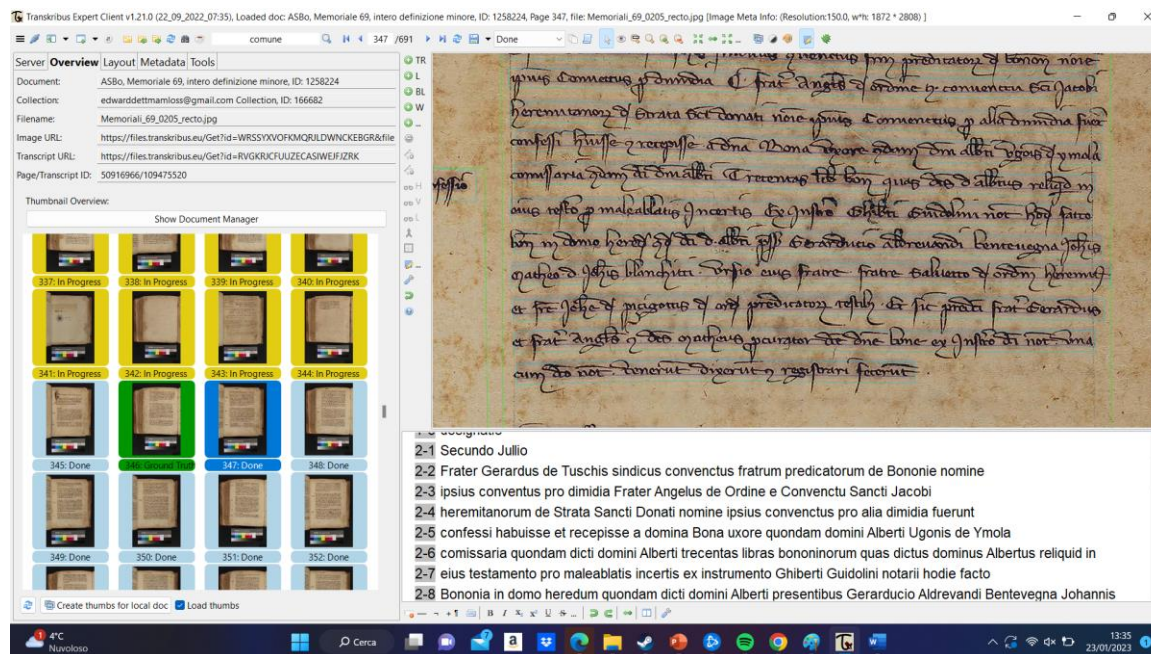
Put aside in the initial test

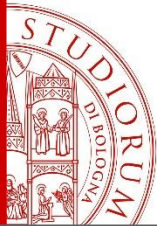
MemoBo and Transkribus

Models of data input for the creation of «ground truth»

2) Direct transcription

- Expansion of abbreviations directly in the main text





MemoBo models of data input: Direct transcription

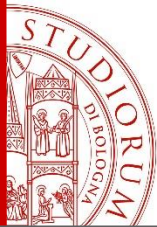
Advantages:

- **Quicker to produce ground truth:**
estimated 10 minutes a page
- No need to use external characters and fonts (avoid subsequent errors)

Disadvantages:

- **Larger amount of material necessary to produce first model**
- **Necessity of more experienced and prepared trainers (paleography and latin)**
- **Long preparatory phases (decisions on how to transcribe vowels and consants, on how to expand abbreviations and etc.**

Chosen for the initial test



MemoBo, Me. 69, Model 1

General description:

Data input method: direct transcription

Some of the transcription specifics:

«U» transcribed as «v» in between vowels.
«j» kept as «j» in the beginning of sentences, but transcribed as «i» between consants

Capitals introduced for names, surnames, patronymic and locations

Decorations and large capitals excluded from the general «layout» - too few to explore in the model.

Aim: try to establish a functioning model with the least amount of material (ratio between effort and result)

Numbers of pages transcribed: 15

Number of **lines**: 451

Number of **words**: 4981

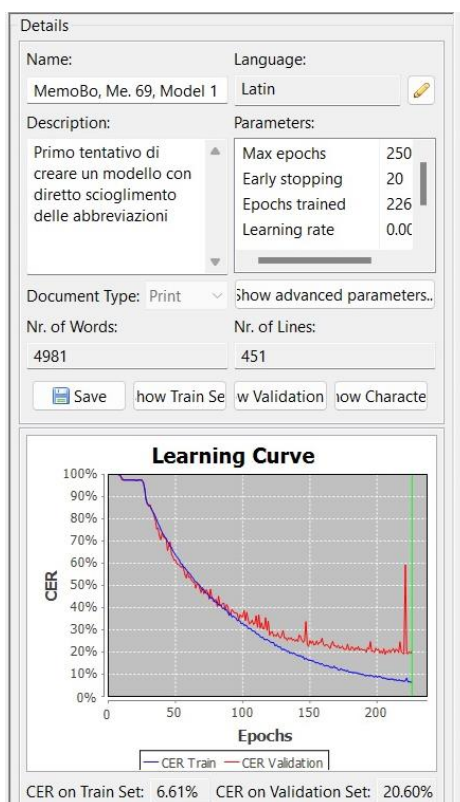
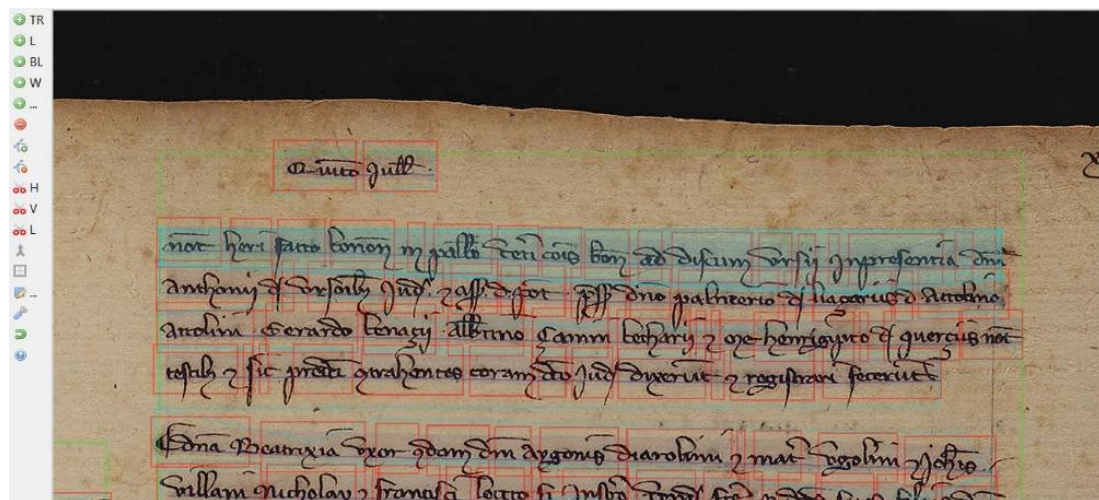
Number of **hours**: 3 hours

MemoBo, Me. 69, Model 1

Results:

CER on Train Set: 6,61%

CER of Validation Set: 20,60% (poor) – ideal: under 10%

2-1 Quinto Julii

2-2 notarii hei facto Bononie in pallio veteri comunis Bononie ad discum Ux in io presentia domi

2-3 Anthoy de Usoibus iudicis instois domini potesatis presentibus domino Palbmerio de Lugnos domini Narobono

2-4 artolini Gerardo Beonaqii Albertino cani Bachoi et me Henrigipto de Querqis notaie

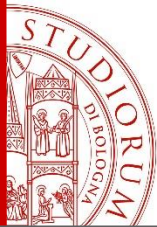
2-5 testibus et si predicti conntrahentes toram dicto iudicis dixerunt et registrara fecerunt

2-6 Domina Roarxia uxor quondam domini Aiginis Darolini et nomter Ugolini J-oJohannes

2-7 vilani Nicholay et Frmes cini loco si instrumento vendione facte pro dictus sus sfuus domini

2-8 Vabeio

In practical terms: 8 mistakes every 15 words



MemoBo, Me. 69, Model 2

General description:

Data input method and transcription specifics: same as *MemoBo, Me. 69, Model 1*

Numbers of pages transcribed:
25

Number of **lines**: 729

Number of **words**: 8149

Number of **hours**: 5 hours

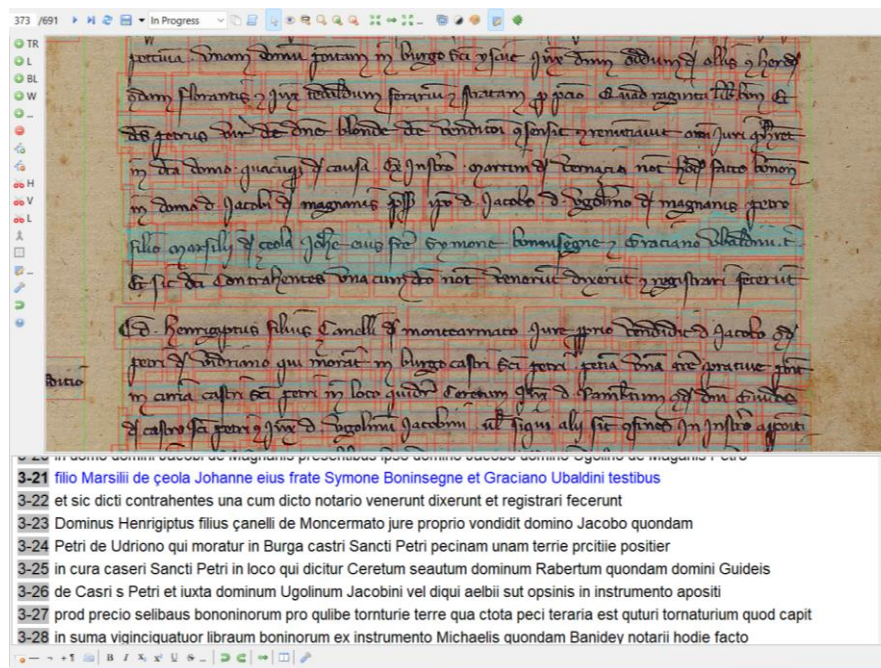
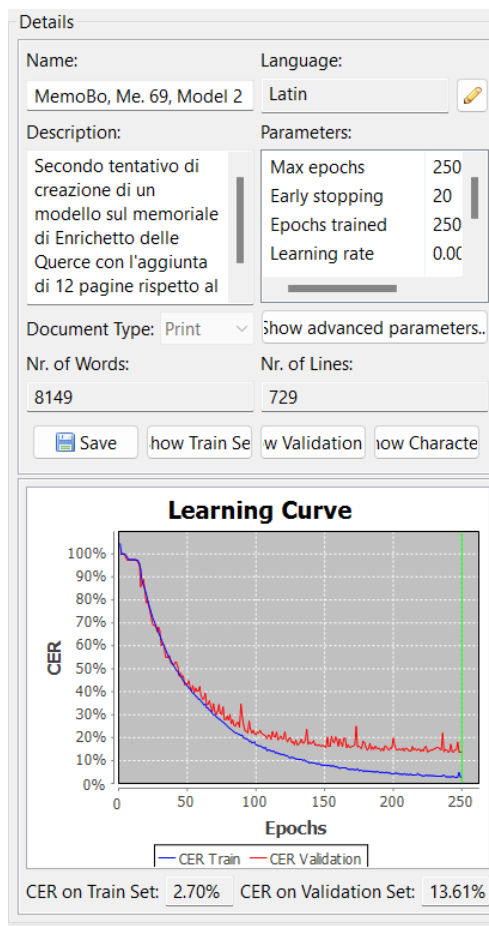
Dataset Overview				×
Train Set:				
Data Type	Pages	Lines	Wor...	
Document ...	24	739	8256	
Ground Trut...	0	0	0	
Total	24	739	8256	
Validation Set:				
Data Type	Pages	Lines	Wor...	
Document ...	3	96	1010	
Ground Trut...	0	0	0	
Total	3	96	1010	
Start Training				Cancel

MemoBo, Me. 69, Model 2

Results:

CER on Train Set: 2,70% (reduced in 60%)

CER on Validation Set: 13,61% (reduced in 34%)



Practical terms: 5 mistakes every 15 words, with some perfect lines.



MemoBo, Me. 69, Model 3

General description:

Data input method and transcription specifics: same as *MemoBo, Me. 69, Model 1* and *MemoBo, Me. 69, Model 2*

Numbers of pages transcribed: 40

Number of **lines**: 1074

Number of **words**:

Number of **hours**: 8 hours

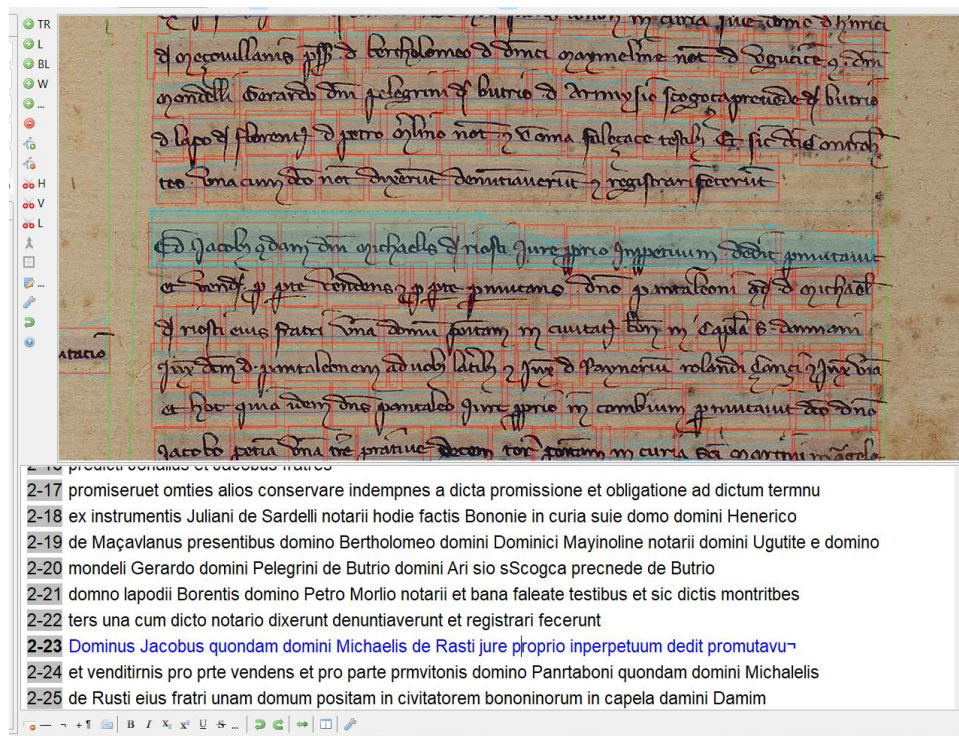
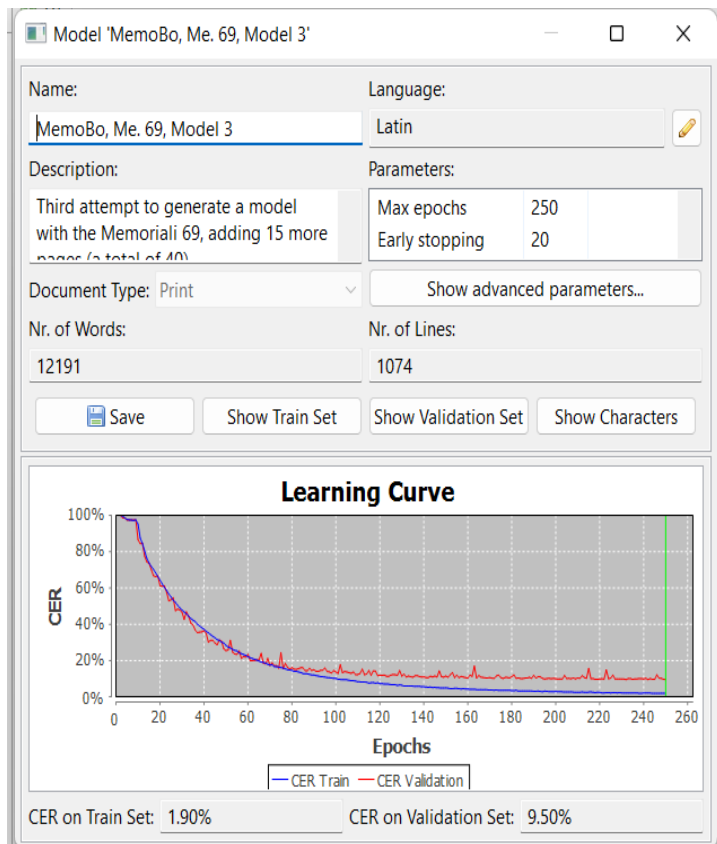
Dataset Overview				×
Train Set:				
Data Type	Pages	Lines	Wor...	
Document ...	35	1074	121...	
Ground Trut...	0	0	0	
Total	35	1074	12...	
Validation Set:				
Data Type	Pages	Lines	Wor...	
Document ...	5	161	1761	
Ground Trut...	0	0	0	
Total	5	161	1761	
<input type="button" value="Start Training"/>				<input type="button" value="Cancel"/>

MemoBo, Me. 69, Model 3

Results

CER on Train Set: 1,90% (reduced in 30%)

CER on Validation Set: 9,50% (reduced in 30%)



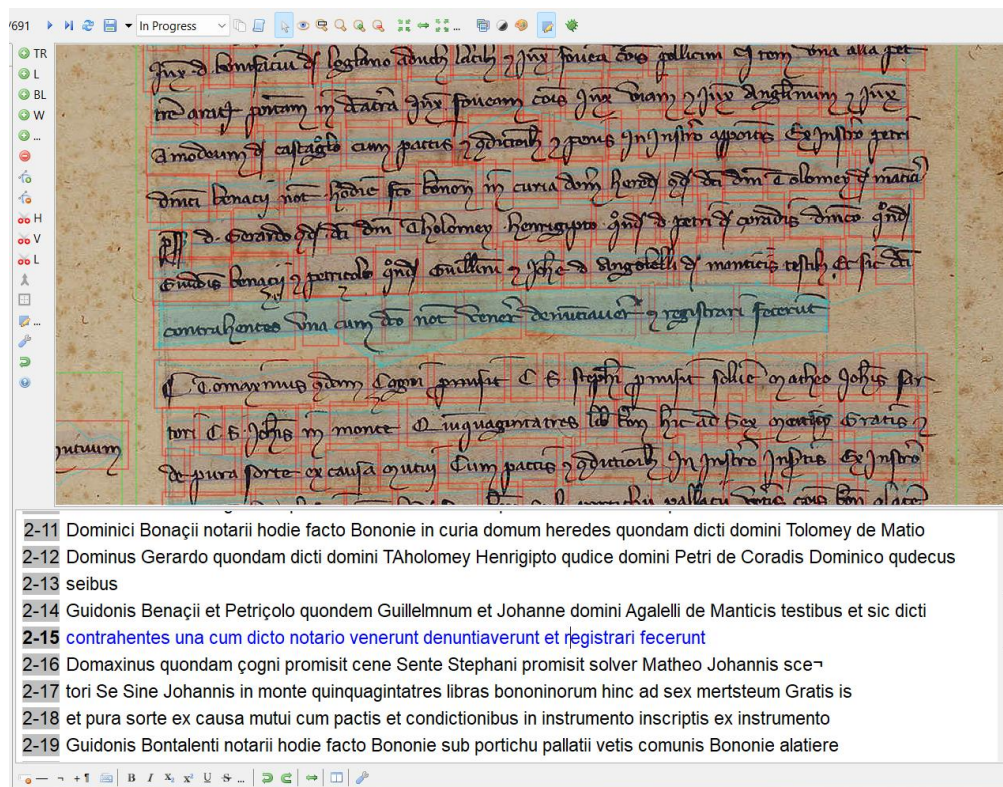
Practical terms: 2-3 mistakes every 15 words, with an average of perfect lines per page.

MemoBo, Me. 69, Model 3

According to *Transkribus*, *MemoBo, Me. 69, Model 3* could be considered an already optimal model: CER on validation under 10%

In diplomatic terms, the model is based on only 20 *folios* (*recto* and *verso*) out of the 300 that compose the entire archival unit = less than 10% of the entire unit allowed us to create an effective model!

Es.) Model MemoBo, Me. 69, Model 3 applied on a page completely not previously known to the software.



691 In Progress

TR
L
BL
W
...
H
V
L
...
...

2-11 Dominici Bonaçii notarii hodie facto Bononie in curia domum heredes quondam dicti domini Tolomey de Matio
2-12 Dominus Gerardo quondam dicti domini TAholomey Henrigo quodice domini Petri de Coradis Dominico quodcus
2-13 seibus
2-14 Guidonis Benaçii et Petriçolo quondem Guillelmnum et Johanne domini Agalelli de Manticiis testibus et sic dicti
2-15 **contrahentes una cum dicto notario venerunt denuntiaverunt et registrar fecerunt**
2-16 Domaxinus çogni promisit cene Sente Stephani promisit solver Matheo Johannis sce-
2-17 tori Se Sine Johannis in monte quinquagintatres libras bononinorum hinc ad sex mertsteum Gratis is
2-18 et pura sorte ex causa mutui cum pactis et conditionibus in instrumento inscriptis ex instrumento
2-19 Guidonis Bontalenti notarii hodie facto Bononie sub portichu pallatii vetis comunis Bononie alatiere

Does the model work for other units of the *Memoriali* series ?

First attempt

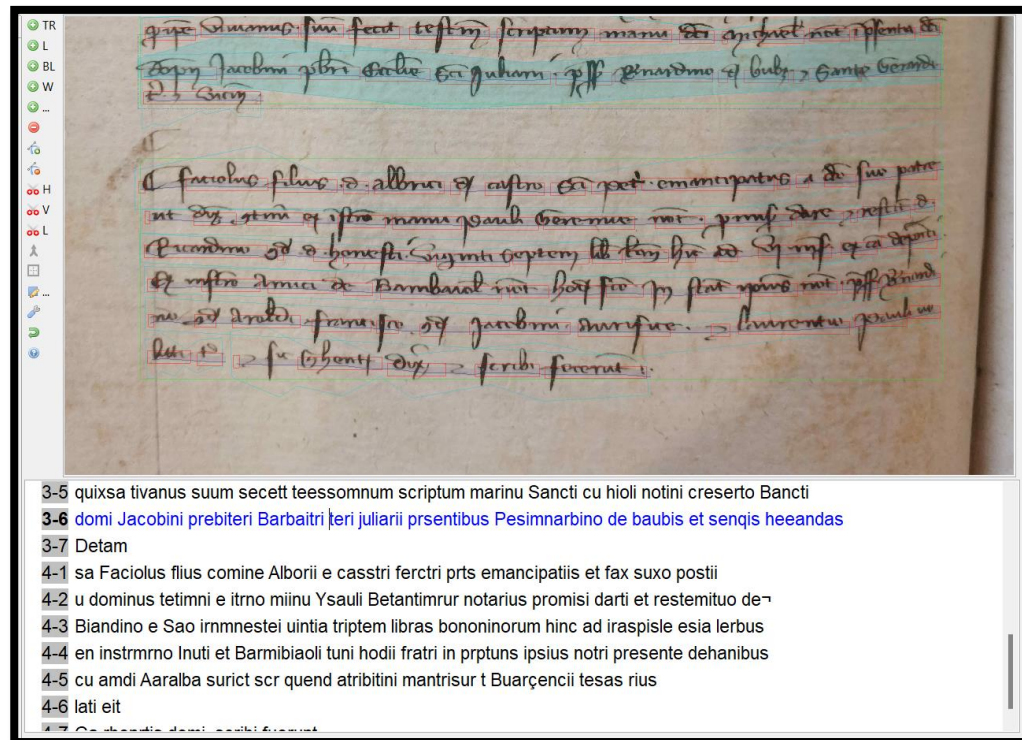
Memoriale of Nicola from Lastignano (1286) – only one year earlier than Enrichetto' unit

Same abbreviation system

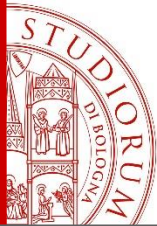
Same script style

Similar clarity of support and similar gramatic consistency

Identical cutural background

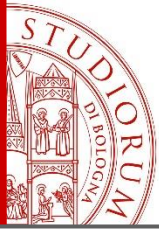


- Yes...but, not really.
- The model is able to recognise many correct characters but there are mistakes in almost every word.



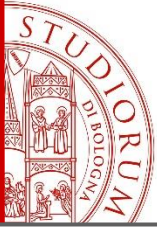
Some (partial) conclusions of the work in progress:

- *Transkribus* is very sensitive to the change of hands even inside the same volume of the *memoriali*
- The software finds it very difficult to transcribe words it has never seen (necessity to augment the sample)
- The homogeneity of the script seems to be more important than the the abbreviation system (the software learnt quickly how to associate abbreviations with the expanded words in the direct transcription)



Future steps

- Create models with mixed «ground truth» in equally representative parts: larger sample.
- Accept the necessity for multiple models: pay close attention to the ratio of time vs cost vs benefit



Questions?

edward.loss2@unibo.it