Comparing human and machine performances in transcribing 18th century handwritten Venetian script

Sofia Ares Oliveira

Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne

DH2018, June 2018, Mexico City

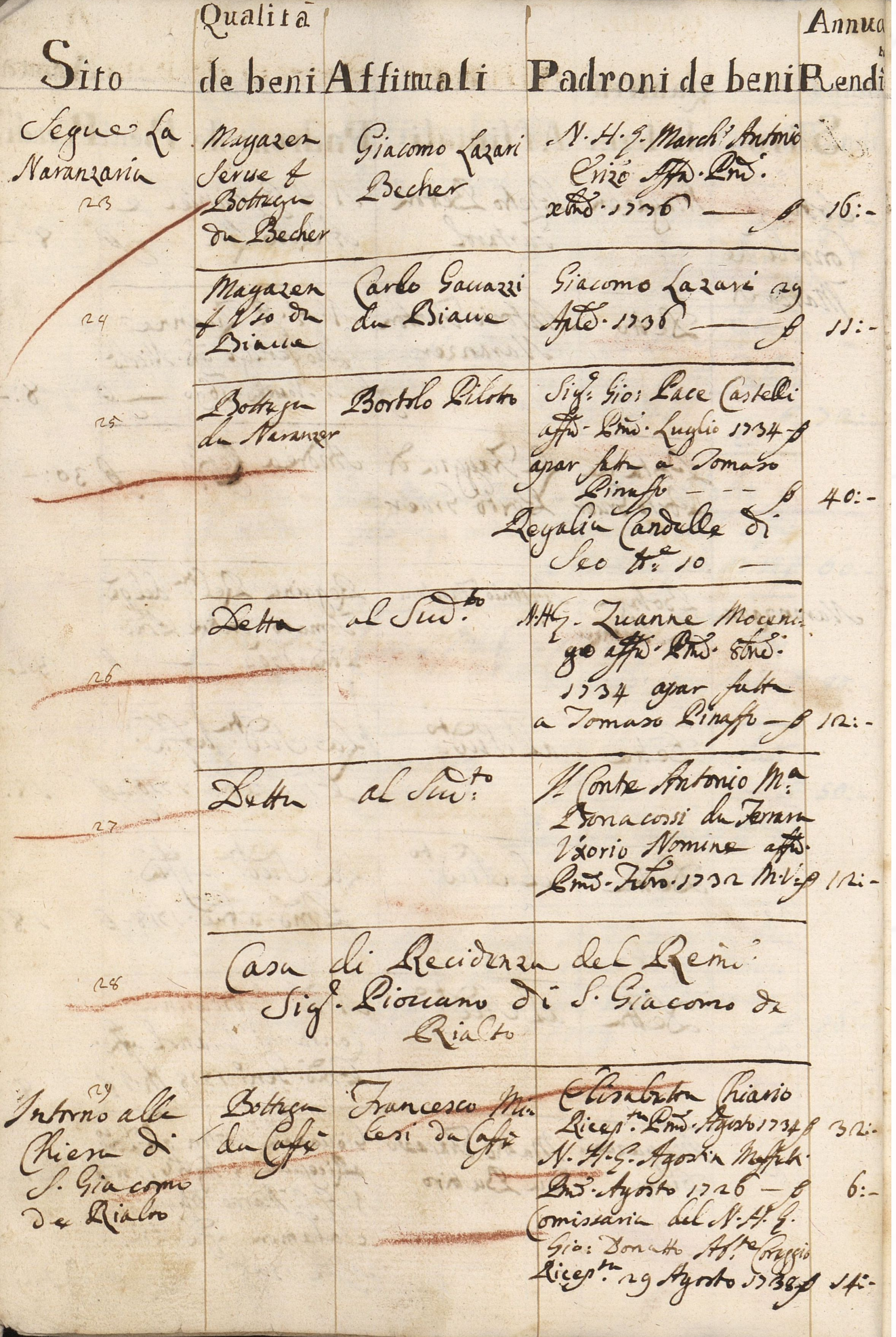
Objective :

Make collections of digitized archival records accessible through textual search

Usually two options for transcription of large collections :

- Crowdsourcing
- Automatization

Can automatic transcription be as good as crowdsourcing approach ?

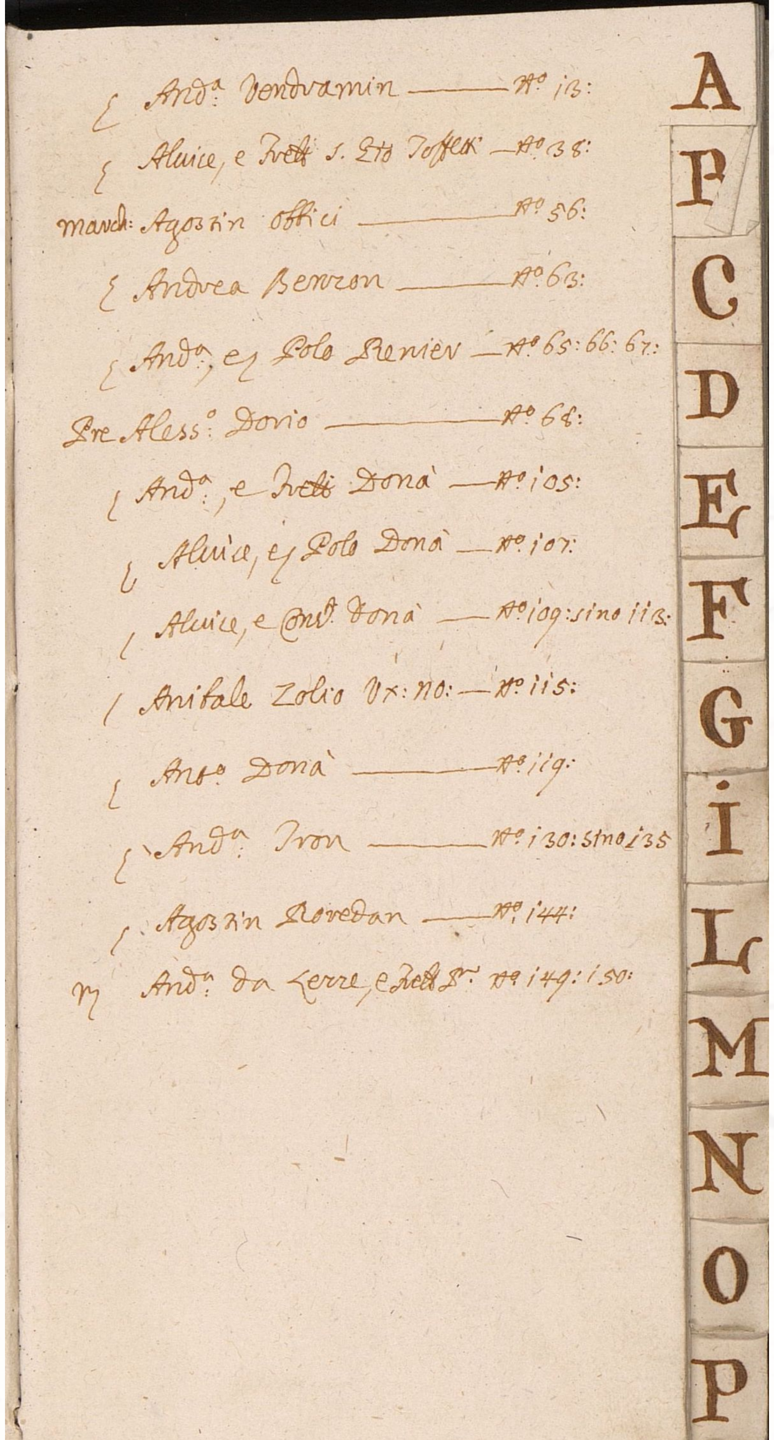


Sito	Qualità de beni	Affituali	Padroni de beni	Annua Rendi
Segue La Naranzaria 23	Magazen Serve & Bottega da Becher	Giacomo Lazari Becher	N.H.G. Marchº Antonio Erizo Affº Crñ. Xbre 1736 — ₤	16:-
24	Magazen & Uso da Biava	Carlo Gavazzi da Biave	Giacomo Lazari 29 Aple 1736 — ₤	11:-
25	Bottega da Naranzer	Bortolo Piloto	Sigr Gio: Pace Castelli affº Crñ. Luglio 1734 ₤ apar fatta a Tomaso Linasso — — ₤ Regalia Candelle di Seo Nº 10 —	40:-
26	Detta	al Sudto	N.H.G. Zuanne Mocenigo affº Crñ. 8bre 1734 apar fatta a Tomaso Linasso — ₤	12:-
27	Detta	al Sudto	Il Conte Antonio Mª Bonacossi da Ferrara Uxorio Nomine affº Crñ. Febrº 1732 M.V. ₤	12:-
28	Casa di Recidenza del Remº Sigr Pievano di S. Giacomo da Rialto			
29 Intorno alla Chiesa di S. Giacomo da Rialto	Bottega da Caffè	Francesco Maresi da Caffè	Elisabetta Chiario Licepta Crñ. Agosto 1734 ₤ N.H.G. Agostin Moffetti Crñ. Agosto 1726 — ₤ Comissaria del N.H.G. Gio: Donatto Asse Correggio Licepta 29 Agosto 1738 ₤	32:- 6:- 14:-

Dataset

Subset of 18th century fiscal documents from the Venetian State Archives

Catastici delle parrocchie



Andª Vendramin	Nº 13:
Aluise, e Frelli s. Sto Toffeti	Nº 38:
march: Agostin Obbici	Nº 56:
Andrea Benzon	Nº 63:
Andª, e Polo Renier	Nº 65: 66: 67:
Pre Alessº Dorio	Nº 68:
Andª, e Frelli Donà	Nº 105:
Aluise, e Polo Donà	Nº 107:
Aluise, e Contº Donà	Nº 109: sino 113:
Anibale Zolio Ux: no:	Nº 115:
Antº Donà	Nº 119:
Andª Tron	Nº 130: sino 135
Agostin Loredan	Nº 144:
m Andª da Lezze, e Frelli Sri	Nº 149: 150:

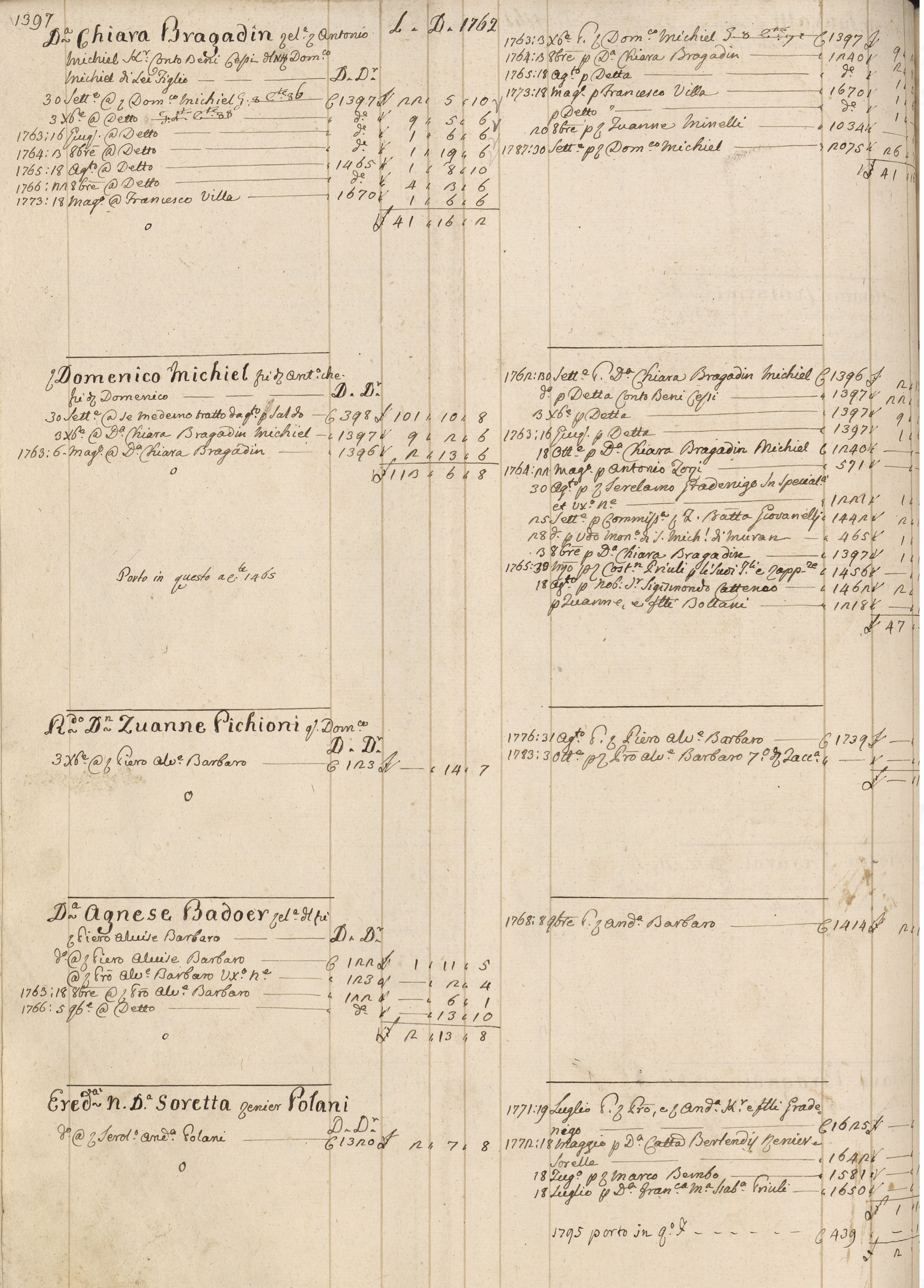
A B C D E F G I L M N O P

Dataset

Subset of 18th century fiscal documents from the Venetian State Archives

Catastici delle parrocchie

Indici



1397
Dª Chiara Bragadin relª q Antonio Michiel K.r Conto Beni Cespi dell'q Domco Michiel di Lei Figlio — L. D. 1762 — D. D.r
30 Sett.e @ q Domco Michiel q. Cte 36 — 6 1397 £ 22 5 10
3 9bre @ Detto — d.o 9 5 6
1763:16 Lugl. @ Detto — d.o 1 6 6
1764:13 8bre @ Detto — d.o 1 19 6
1765:18 Agto @ Detto — 1465 1 8 10
1766:22 8bre @ Detto — d.o 4 3 6
1773:18 Magg.o @ Francesco Villa — 1670 1 6 6
£ 41 16 2

1763:3 9bre ... q Domco Michiel — 6 1397 £
1764:13 8bre p Dª Chiara Bragadin — 1240
1765:18 Agto p Detta — d.o
1773:18 Magg.o p Francesco Villa — 1670
p Detto — 1034
20 8bre p q Zuanne Minelli — 1034
1787:30 Sett.e p q Domco Michiel — 2075 26
£ 41

q Domenico Michiel fu q Ant.o che fu q Domenico — D. D.r
30 Sett.e @ se medesimo tratto dagli p Saldo — 6 398 £ 101 10 8
3 9bre @ Dª Chiara Bragadin Michiel — 1397 9 2 6
1763:6 Magg.o @ Dª Chiara Bragadin — 1396 2 13 6
£ 113 6 8
Porto in questo ac.te 1465

1762:30 Sett.e ... Dª Chiara Bragadin Michiel — 6 1396 £
d.o p Detta Conto Beni Cespi — 1397
3 9bre p Detta — 1397
1763:16 Lugl. p Detta — 1397
18 Ott.e p Dª Chiara Bragadin Michiel — 6 1240
1764:22 Magg.o p Antonio Zorzi — 571
30 Agto p q Serelamo Fradenigo In Special. et Ux.r N.e — 1222
25 Sett.e p Commiss.a q Z. Batta Giovanelli — 1442
28 d.o p Uso mon.o di S. Mich.l di Murano — 465
3 8bre p Dª Chiara Bragadin — 1397
1765:30 Mag.o p Cost.n Friuli ... — 1456
18 Agto p Nob. S.r Sigismondo Cattaneo — 1462
p Zuanne e f.lli Bollani — 1218
£ 47

N.ro D.n Zuanne Pichioni q Domco — D. D.r
3 9bre @ q Piero Alv.e Barbaro — 6 123 £ 14 7

1776:31 Agto ... q Piero Alv.e Barbaro — 6 1739 £
1783:3 Ott.e p q Pro Alv.e Barbaro 7.o q Zacc.a

Dª Agnese Badoer relª q Piero q Piero Aluise Barbaro — D. D.r
d.o @ q Piero Aluise Barbaro — 6 122 £ 1 11 5
@ q Pro Alv.e Barbaro Ux.r N.e — 123 22 4
1763:18 8bre @ q Pro Alv.e Barbaro — 122 6 1
1766:5 9bre @ Detto — d.o 13 10
£ 2 13 8

1768:8 9bre q Andr.a Barbaro — 6 1414

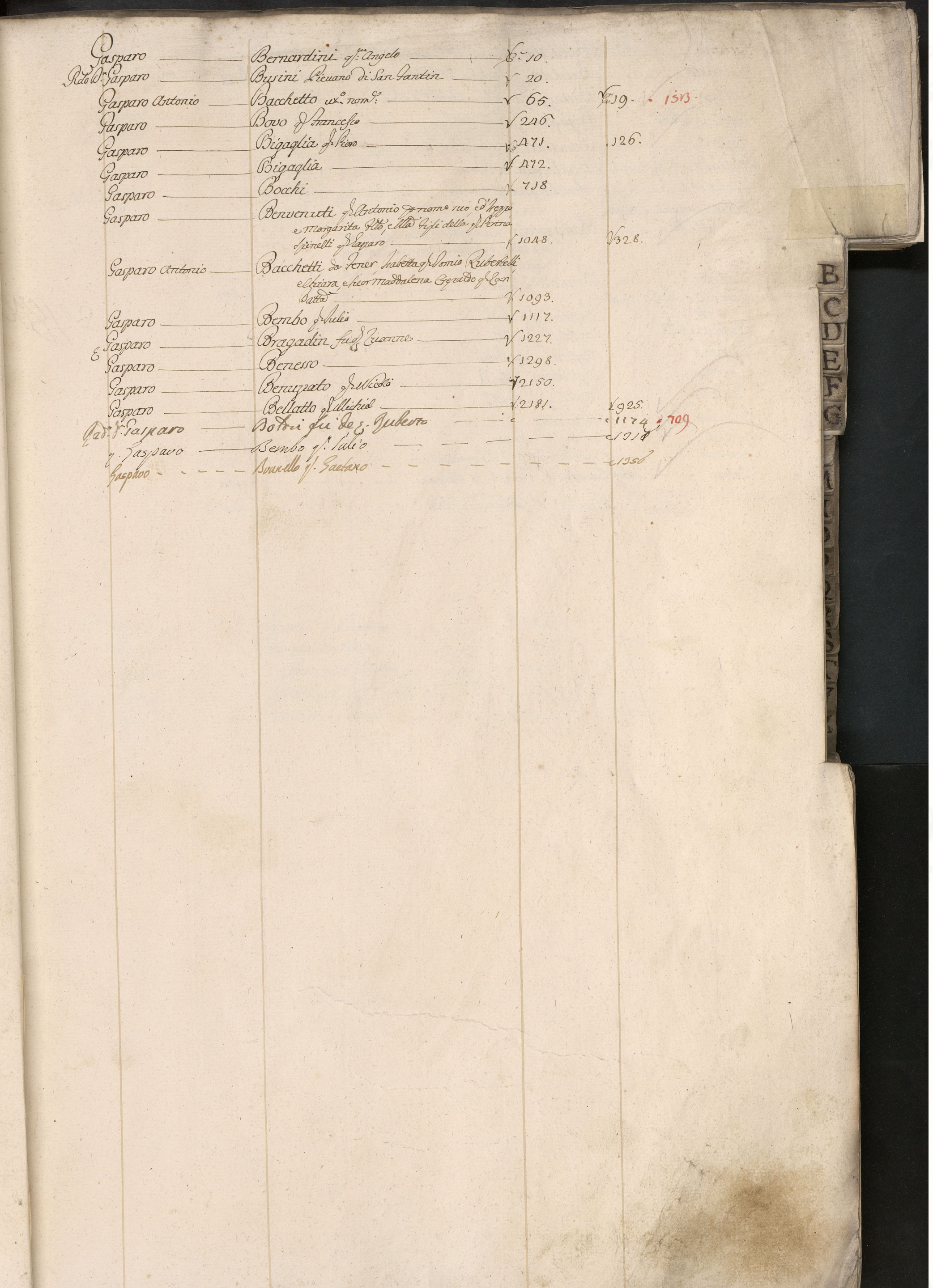
Ered.i N.D. Soretta Renier Polani — D. D.r
d.o @ q Zerol.o Andr.a Polani — 6 1320 £ 2 7 8

1771:19 Luglio q Pro, e q Andr.a K.r e f.lli Gradenigo — 6 1625
1772:18 Maggio p Dª Catta Berlendis Renier Sorelle — 1642
18 Lug.o p q Marco Bembo — 1581
18 Luglio p Dª Fran.ca M.a Isab.a Friuli — 1650
1795 porto in q.o £ — 6 439

Dataset

Subset of 18th century fiscal documents from the Venetian State Archives

- Catastici delle parrocchie
  - Indici
- Quaderni dei Trasporti

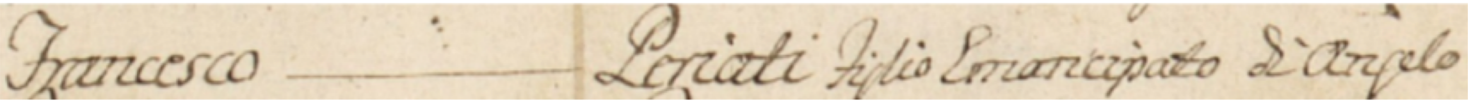


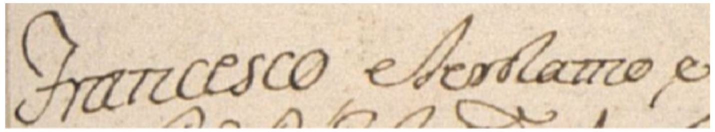
Gasparo	Bernardini q.m Angelo	c. 10.	
Rdo D.r Gasparo	Busini Pievano di San Fantin	√ 20.	
Gasparo Antonio	Bacchetto ux.o nom.e	√ 65.	√19. · 1513
Gasparo	Bovo q. Francesco	√ 246.	
Gasparo	Bigaglia q. Piero	√ 471.	126.
Gasparo	Bigaglia	√ 472.	
Gasparo	Bocchi	√ 718	
Gasparo	Benvenuti q. Antonio per nome suo ed Iseppo e Margarita Fr.lli, e Sorelle figli della q. Perina Spinelli q. Gasparo	√ 1048.	√328.
Gasparo Antonio	Bacchetti da Fener, Isabetta q. Domenico Rubertelli e Chiara, e suor Maddalena Geraldo q. Zan Batta	√ 1093.	
Gasparo	Bembo q. Iulio	√ 1117.	
E Gasparo	Bragadin fu q. Zuanne	√ 1227.	
Gasparo	Benesso	√ 1298.	
Gasparo	Beruzzato q. Nicolò	√ 2150.	
Gasparo	Bellatto q. Michiel	√ 2181.	√925.
Rdo P. Gasparo	Bottari fu de q. Ruberto		1174 · 709
q. Gasparo	Bembo q. Iulio		1318
Gasparo	Bonvello q. Gaetano		1350

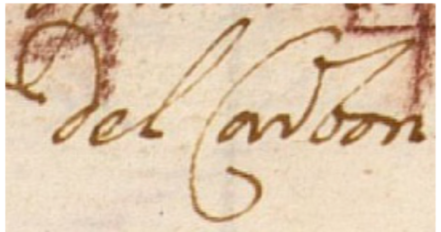
Dataset

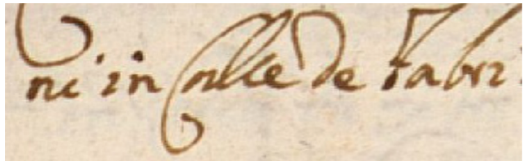
Subset of 18th century fiscal documents from the Venetian State Archives

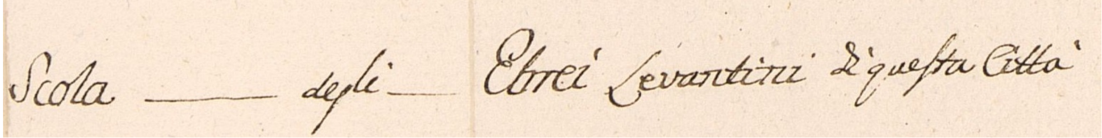
- Catastici delle parrocchie
  - Indici
- Quaderni dei Trasporti
  - Indici

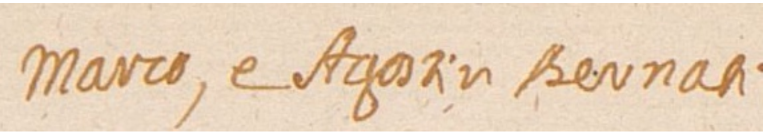

Francesco —— Peziati figlio Emancipato di Angelo

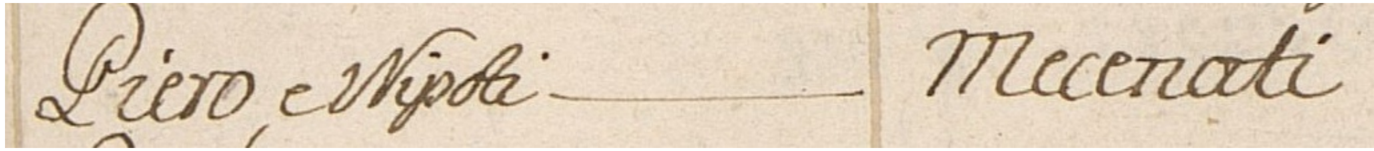

Francesco e Bortolamio e

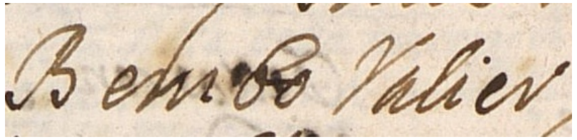

del Cardon

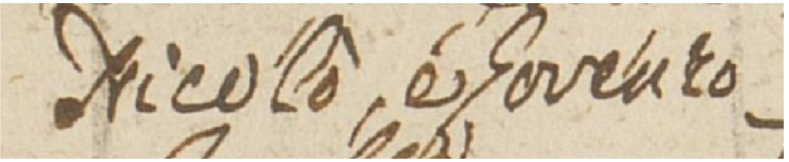

ni in Calle de Fabri

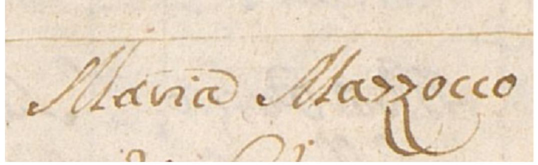

Scola —— degli —— Ebrei Levantini di questa Città

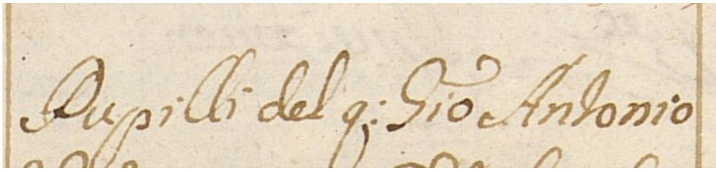

Marco, e Agostin Bernar

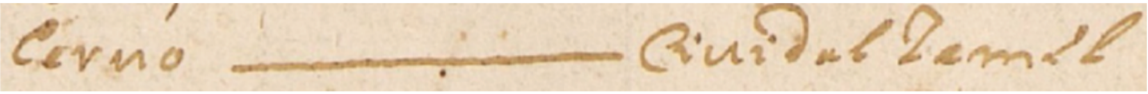

Piero, e Nipoti —— Mecenati

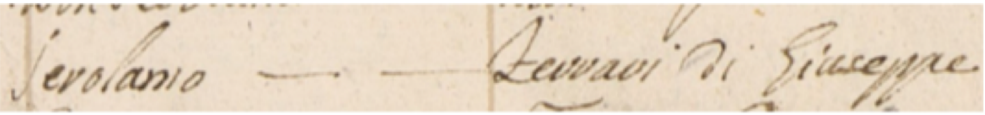

Bembo Valier

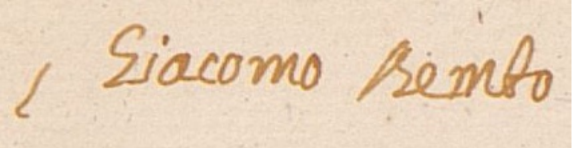

Nicolò, e Zovenuto

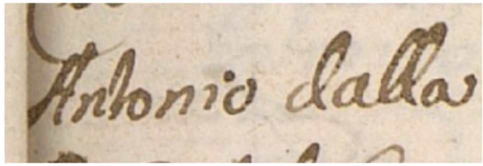

Maria Mazzocco

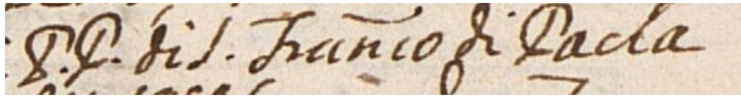

Pupilli del q. Gio. Antonio

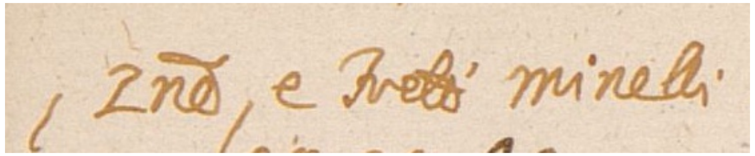

Cernò —— Civil dal Zamel

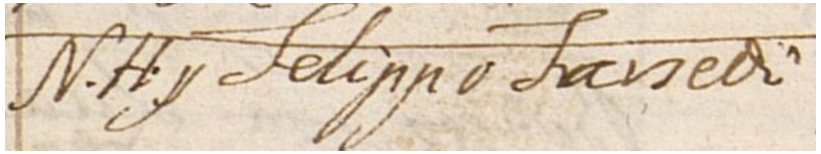

Gerolamo —— Zennari di Giuseppe

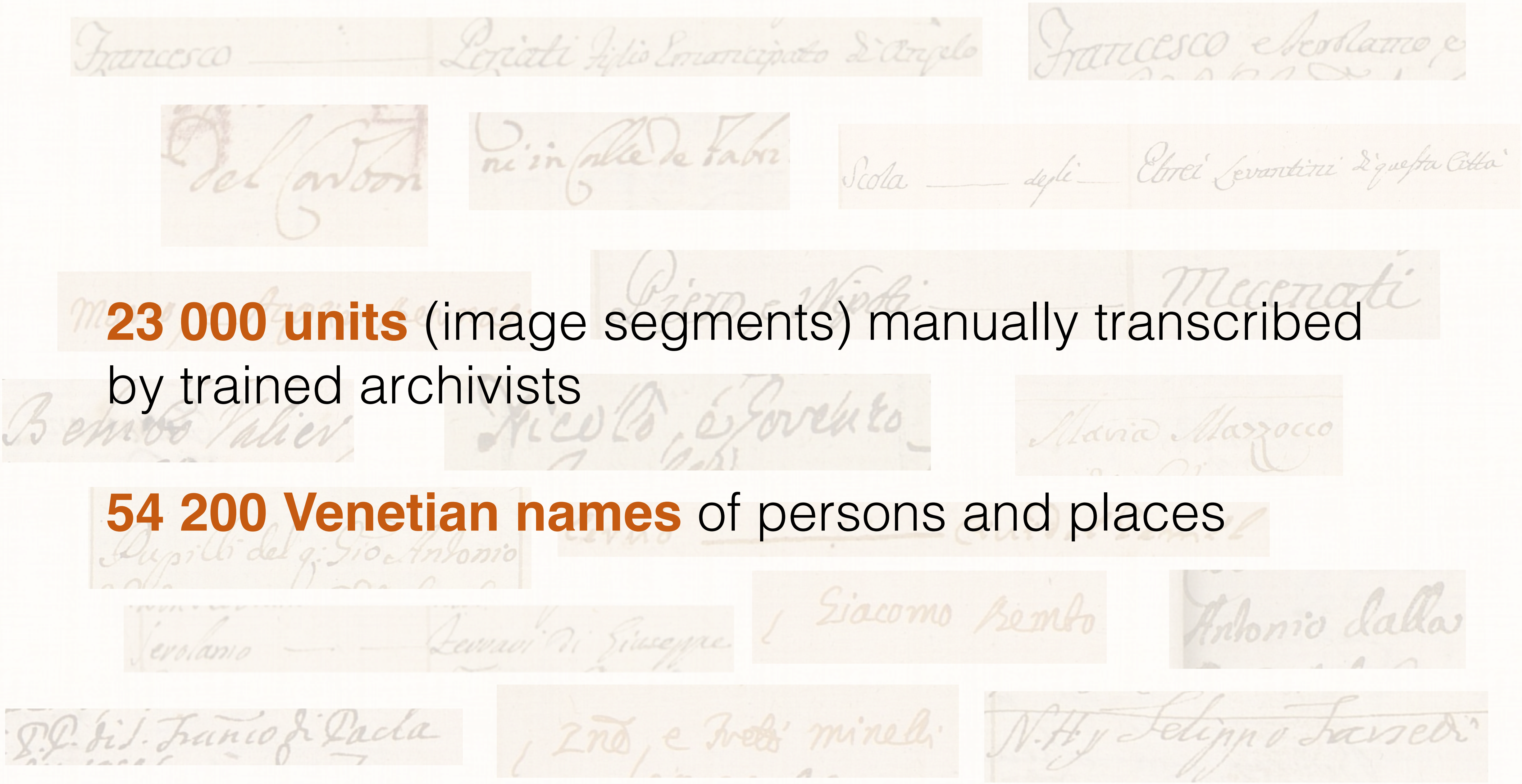

(Giacomo Bembo


Antonio dalla


R.R. di S. Franco di Paola


(Zne, e Frelli mineli


N.H. Filippo Farsetti


Francesco
Scola degli Ebrei Levantini di questa Città
Mecenati
Maria Marzocco
Giacomo Bembo
Antonio dalla

23 000 units (image segments) manually transcribed by trained archivists

54 200 Venetian names of persons and places

Machine transcription

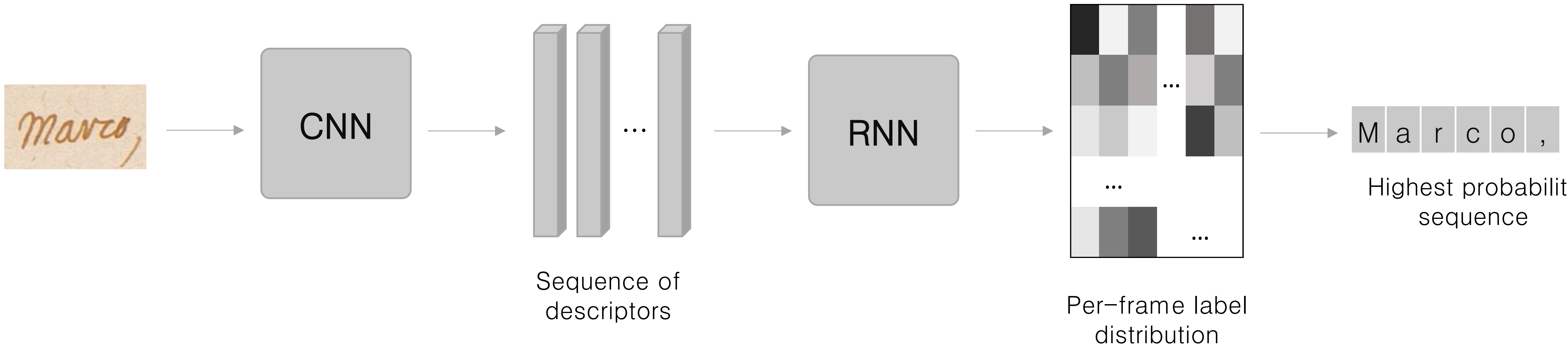
vs

Human transcription

http://home.bt.com/images/turing-bombe-136399065494703901-150706144229.jpg

Neural network architecture : CRNN

A Convolutional Recurrent Neural Network combines the best of convolutional and recurrent neural networks



B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” 2017

Training

Height is fixed for all the image segments (but image ratio is kept)

Data augmentation (contrast, intensity, rotation)

20 712 image segments

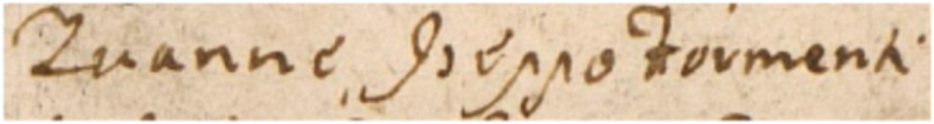
48 628 words in total

8 848 vocabulary items

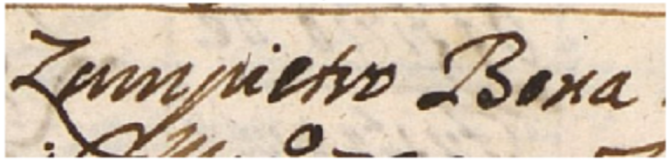
‘A-Za-z’ characters + a few symbols for punctuation

Evaluation

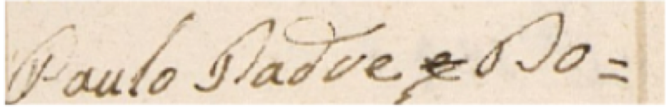
P: Zuanne Iseppo Formenti
GT: Zuanne, Iseppo Formenti



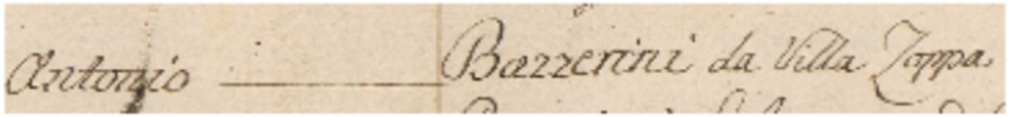
P: Zuanpietro Bona
GT: Zampietro Bona



P: Paulo Padre, e Do=
GT: Paulo Padre, Do=



P: Antonio Bazzerini da Villa Zappa
GT: Antonio Bazzerini da Villa Zoppa



2 317 image segments
5 559 words in total
2 157 vocabulary items

Character Error Rate (CER) : **0.0804**

Machine transcription

vs

Human transcription

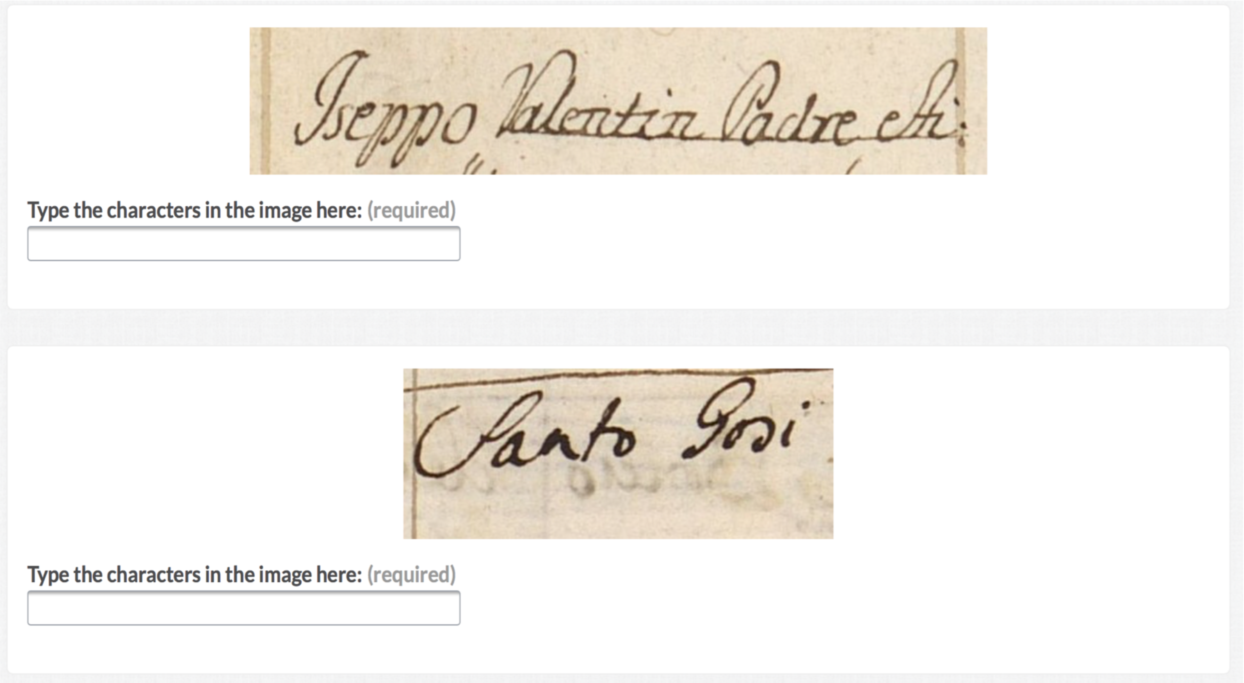
http://footage.framepool.com/shotimg/qf/775872743-archivist-historians-librarian-manuscript.jpg

Evaluation of the average performance of Italian-speaking transcribers

Platform : CrowdFlower (now Figure Eight)

Task : Transcribe text in image segment, taking into account capitals and punctuation

Data : 2 317 image segments from test set



Type the characters in the image here: (required)

Type the characters in the image here: (required)

Experiment

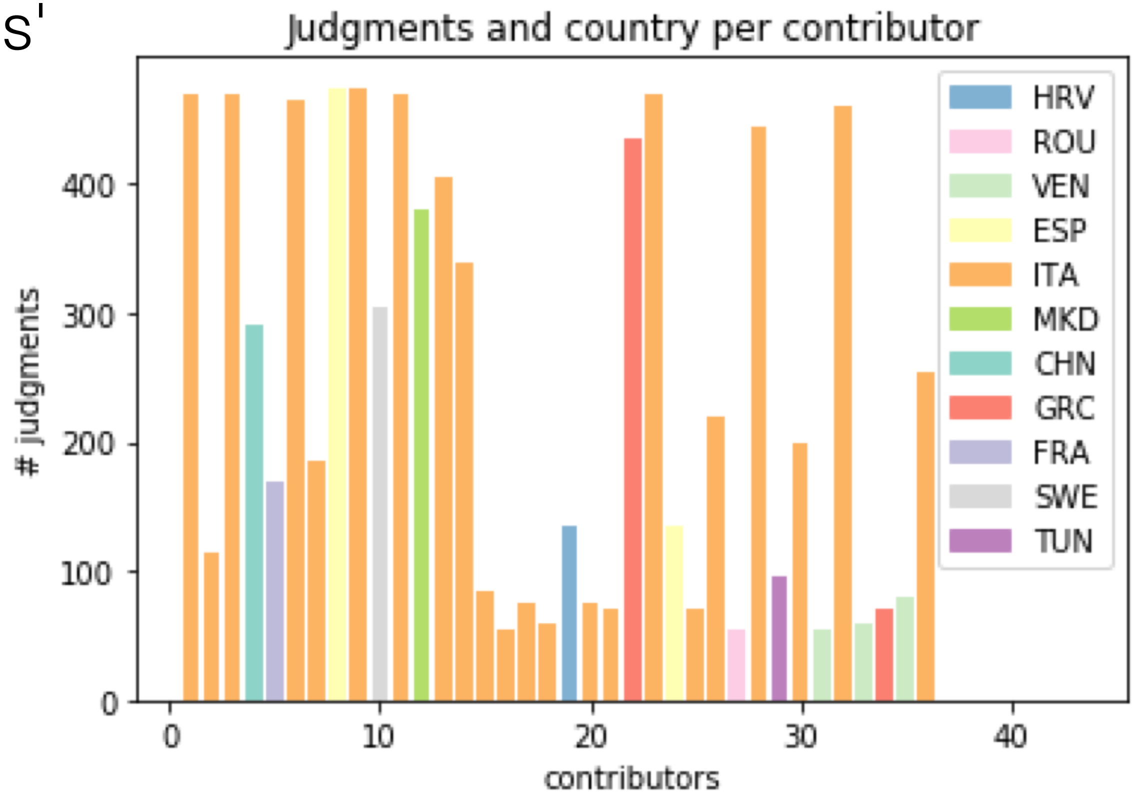
Evaluation of the reliability of transcribers' answers during the experiment :

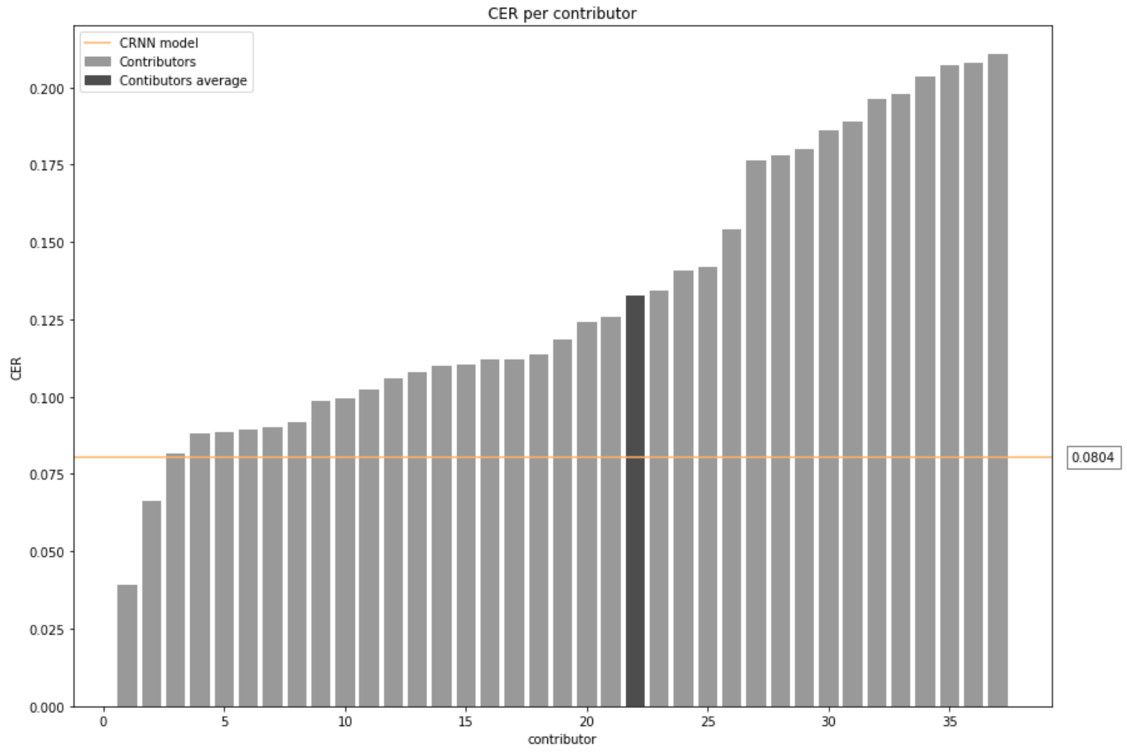
103 evaluation units

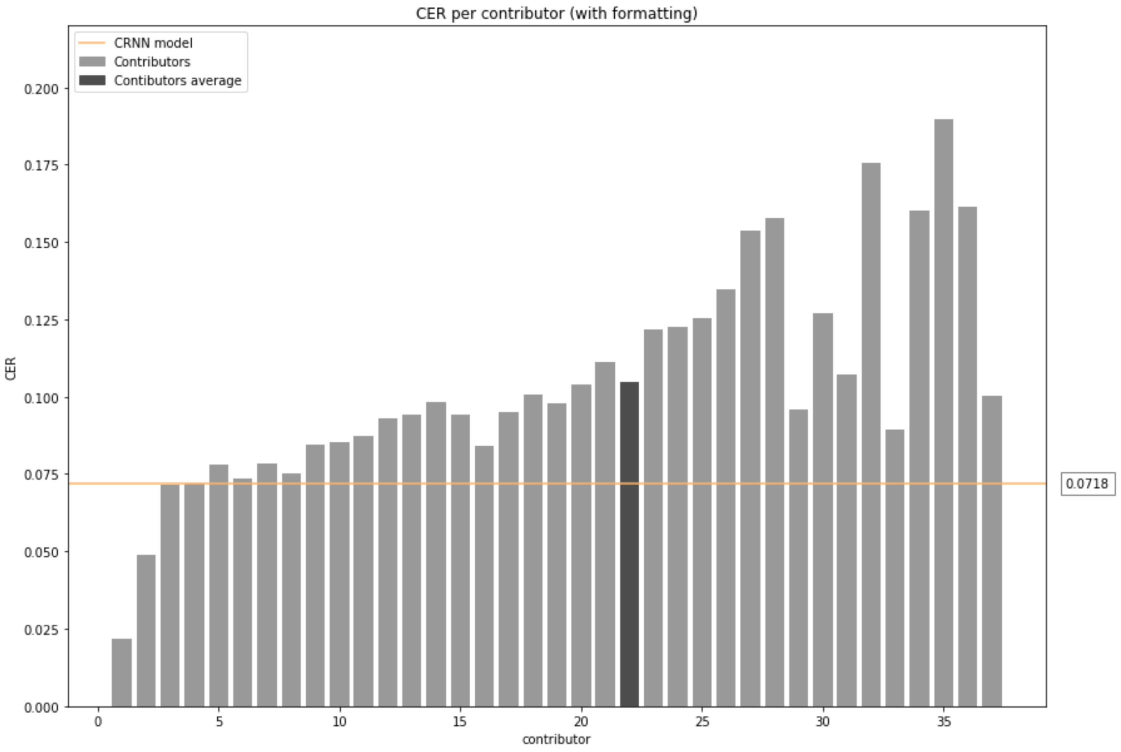
0.6 accuracy required

36 transcribers remained after selection

8 674 valid transcriptions to analyze







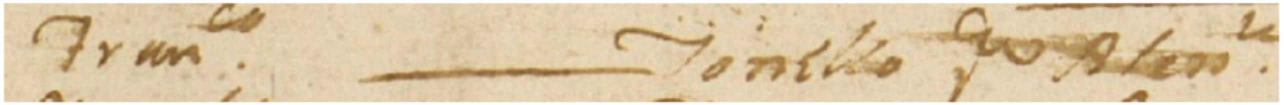
	CER system	CER amateur	WER system	WER amateur
No formatting	0.0804	0.1328	0.2709	0.4318
With formatting*	0.0718	0.1047	0.2551	0.3507

*formatting = replace capital letters by lowercase and remove punctuation

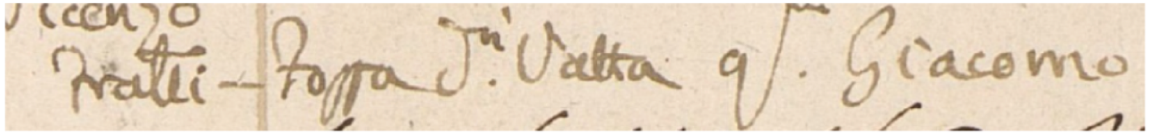
On going work

Extend the recognition to abbreviation symbols

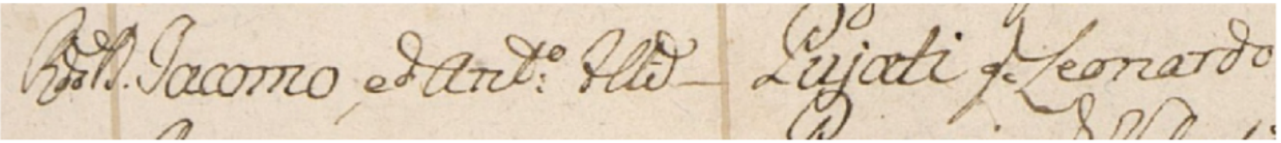
P: francesco Zonello quondam alessandro
GT: francesco Tonello quondam alessandro



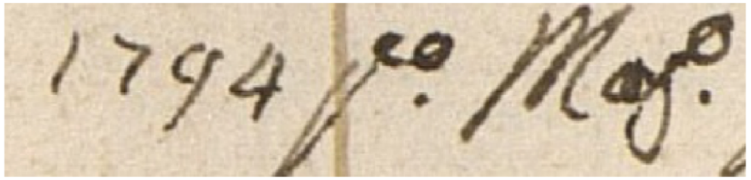
P: fratelli Fappa detta Vattin quondam Giacomo
GT: fratelli Foppa detti Vatta quondam Giacomo



P: reverendo don Iacomo, ed antonio fratelli Picati quondam Leonardo
GT: reverendo don Iacomo, ed antonio fratelli Pujati quondam Leonardo



P: 1794 primo maggio
GT: 1794 primo maggio



The system has lower CER and WER than amateur transcribers' average on 18th century Venetian script

→ Sufficiently reliable to use for searching purposes

→ New prospects for analyzing and study large collections of documents github.com/solivr/tf-crnn

Venice Time Machine
vtm.epfl.ch

Digital Humanities Laboratory
dhlab.epfl.ch

Sofia Ares Oliveira
sofia.oliveiraares@epfl.ch

Frederic Kaplan
frederic.kaplan@epfl.ch